UNIVERSITY OF COPENHAGEN
DEPARTMENT OF COMPUTER SCIENCE

Ph.D. Thesis

Wenyan Li

Understanding Multimodal Interactions: from Data Variations to Model Behavior

Advisor: Anders Søgaard

This thesis has been submitted to the Ph.D. School of The Faculty of Science, University of Copenhagen on July 8th, 2025.

"The prize is to find things out."
— Richard Feynman

Abstract

Humans develop rich understanding through lifelong sensory experiences and interactions with the world. A child learns what an apple is by seeing its red or green skin, hearing the crisp sound when biting into it, feeling its smooth texture, tasting its sweet flavor, and eventually connecting all these experiences to the word "apple". Just as we effortlessly combine what we see, hear, and touch to understand our surroundings, multimodal systems aim to work with multiple types of information simultaneously, such as image-text pairs, or audio combined with video. In the rapidly evolving landscape of multimodal learning, from describing images with details, to answering challenging questions from visual clues that requires reasoning, this thesis aims to provide an in-depth study of challenges and bottlenecks in building more data efficient, robust and interpretable multimodal models.

Throughout this thesis, we aim to deepen understanding of multimodal learning by focus on three critical areas: how we collect and prepare training data, how we properly evaluate these models within culture contexts, and how we can look inside the complex model components to understand their decision-making processes.

We first investigate data-centric gaps in multimodal learning by introducing dynamic data curation techniques that automatically adjust training datasets based on model status. Through selective sample manipulation, we demonstrate improved image captioning performance without increasing dataset size, effectively addressing the homogeneous treatment of training data that ignores sample difficulty and cross-modal misalignment. To evaluate vision-language models within cultural contexts, we develop FoodieQA, which is a high quality, human-curated benchmark with original images that prevents data contamination in pretrained models. This resource examines regionally-specific food knowledge through multi-image, single-image, and text-only visual question answering tasks, revealing significant limitations in culturally-grounded reasoning among current state-of-the-art models. We then diagnose behavioral blind spots in model representation and predictions. First, in retrieval-augmented captioning models, we identify how retrieved content can mislead predictions and a mitigation strategy based on context sampling. Second, we investigate information loss during modality fusion in vision-language models, demonstrating how projection modules fundamentally alter visual representation geometry. Our novel embedding reconstruction analysis localizes and visualizes specific information degradation at the image patch level, explaining downstream failures in visual question answering and captioning tasks.

From understanding data variations to investigating model behaviors, this thesis delivers insights, practical frameworks, and key datasets required to build the next generation of more reliable, inclusive, and transparent multimodal systems.

Resumé

Mennesker udvikler en rig forståelse gennem livslange sensoriske oplevelser og interaktioner med verden. Et barn lærer, hvad et æble er, ved at se dets røde eller grønne skind, høre den sprøde lyd når der bides i det, føle dets glatte tekstur, smage dets søde smag og til sidst forbinde alle disse oplevelser med ordet "æble". Ligesom vi ubesværet kombinerer, hvad vi ser, hører og rører ved for at forstå vores omgivelser, sigter multimodale systemer mod at arbejde med flere typer information samtidigt, såsom billed-tekst-par eller lyd kombineret med video. I det hurtigt udviklende landskab for multimodal læring, fra at beskrive billeder med detaljer til at besvare udfordrende spørgsmål fra visuelle spor, der kræver ræsonnement, sigter denne afhandling mod at give et dybdegående studie af udfordringer og flaskehalse i opbygningen af mere dataeffektive, robuste og fortolkelige multimodale modeller.

Gennem denne afhandling stræber vi efter at uddybe forståelsen af multimodal læring ved at fokusere på tre kritiske områder: hvordan vi indsamler og forbereder træningsdata, hvordan vi korrekt evaluerer disse modeller inden for kulturelle kontekster, og hvordan vi kan se ind i de komplekse modelkomponenter for at forstå deres beslutningsprocesser.

Vi undersøger først datacentriske huller i multimodal læring ved at introducere dynamiske datakureringsteknikker, der automatisk justerer træningsdatasæt baseret på modelstatus. Gennem selektiv samplemanipulation demonstrerer vi forbedret billedbeskrivelsesydelse uden at øge datasætsstørrelsen, hvilket effektivt adresserer den homogene behandling af træningsdata, der ignorerer samplevanskelighedsgrad og krydsmoral fejljustering. For at evaluere vision-sprog-modeller inden for kulturelle kontekster udvikler vi FoodieQA, som er et benchmark af høj kvalitet, kureret af mennesker med originale billeder, der forhindrer dataforurening i prætrænede modeller. Denne ressource undersøger regionalt specifik madkundskab gennem multi-billed, enkelt-billed og kun-tekst visuelle spørgsmål-svar-opgaver, hvilket afslører betydelige begrænsninger i kulturelt forankret ræsonnement blandt nuværende state-of-the-art modeller. Vi diagnosticerer derefter adfærdsmæssige blinde punkter i modelrepræsentation og forudsigelser. Først identificerer vi i hentningsforstærkede beskrivelsesmodeller, hvordan hentet indhold kan vildlede forudsigelser, og en afhjælpningsstrategi baseret på kontekst-sampling. Dernæst undersøger vi informationstab under modalitetsfusion i vision-sprog-modeller, hvilket demonstrerer, hvordan projektionsmoduler fundamentalt ændrer visuel repræsentationsgeometri. Vores nye rekonstruktionsanalyse af indlejring lokaliserer og visualiserer specifik informationsforringelse på billedpatch-niveau, hvilket forklarer downstream-fejl i visuelle spørgsmål-svar og beskrivelsesopgaver.

Fra forståelse af datavariationer til undersøgelse af modeladfærd leverer denne afhandling indsigt, praktiske rammer og nøgledatasæt, der er nødvendige for at opbygge den næste generation af mere pålidelige, inkluderende og gennemsigtige multimodale systemer.

Acknowledgements

It is unbelievable that this journey has come to an end. The memory of my first day in Copenhagen remains vivid in my mind. As Richard Feynman wisely said, "The prize is to find things out." For me, the three years in PhD is an adventure, which allows me to discover, to satisfy my curiosity, and to enjoy the path from wondering to understanding.

To all my colleagues at CoAStal, former and current, thanks for all the inspiring discussions, support, delightful moments, tea times, ice creams, coffee, lunch, conferences, trip advises, etc. It is you who made CoAStal such a special and lovely place to do a PhD. And I am deeply grateful to my wonderful collaborators and coauthors, without whom the journey would not be as much fun. Working together, our time exploring questions that often seemed far-fetched initially but eventually turned into interesting research has made this experience truly incredible!

I have learned so much from each of the professors at CoAStal—Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. Thank you for engaging with my research ideas, supporting my projects, offering valuable feedbacks, and encouraging me to be a better researcher. Anders, thanks for being there to listen and support during the time that I struggle the most. I extend my thanks to Paul Pu Liang for hosting my visit at MIT and to Ryan Cotterell for welcoming me at ETH Zürich. Special thanks to Jordan Boyd-Graber, my first supervisor in this field; it's been a pleasure to maintain our connection through conferences and to receive your ongoing encouragement for my research endeavors.

I extend my gratitude to Raphael Tang and Ferhan Ture, my former colleagues at Comcast AI, who inspired me and encouraged me to pursue my PhD. Working with you both, on projects within and beyond the company, has been a great pleasure. Special thanks also to Yao Lu and Crystina Zhang. Although we never formally shared a lab, your willingness to discuss any ideas during my PhD and friendship have been invaluable to me.

A PhD journey inevitably includes challenging periods, and I am greatly thankful to my friend Bin Zhang for patiently listening and offering guidance whenever I struggled. I deeply appreciate the help and kindness from Jiaang Li, Yifei Yuan, Ruixiang Cui, Yong Cao, Qiwei Peng, Guimin Hu, Rita Ramos, and Emanuele Bugalio, who supported me through my most difficult moments. To my dear, longtime friends Emily Yue Wang, Yu Zhao, Yajie Mao, Xueyan Li, Xiaomin Lin, and Lei Zheng, thank you for your unwavering support since the beginning, regardless of the physical distance between us.

My experience in Copenhagen would have been entirely different without my friends outside of the lab. Ziyin Li, Yihe Zhang, Bo Cui, Jiahao Lu, Shiwen Yang—your presence made this journey special. And to my cherished friends and best PingPong buddies, Lucas Krieger, Yifan Sun, and Simone Baseggio–you have been an essential part of my PhD life, bringing energy, joy, and balance to my academic pursuits.

Finally, I owe my deepest gratitude to my parents, Junping Li and Juanming Wang, for

your endless support and unconditional love, for always having my back wherever I am, and for respecting each decision I have made to pursue my passion. I am incredibly fortunate to have parents like you, who taught me the value of diligence and optimism. Your support and love have been the foundation of my confidence and perseverance, empowering me to overcome any challenge and pursue everything I strive for. I also want to thank my grandfather, Senlin Wang, who will not be able to witness my graduation but would surely be proud of my academic pursuits and this milestone on my life's journey.

List of publications

This thesis is article-based. The included articles maintain their original content with only minor modifications, such as typo corrections and formatting adjustments for consistency in tables and figures. The articles are presented as chapters in the thesis, in the order listed below:

1. Wenyan Li, Jonas Lotz, Chen Qiu, and Desmond Elliott. The role of data curation in image captioning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1074–1088, St. Julian's, Malta, March 2024d. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.65.

2. Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. FoodieQA: A multimodal dataset for fine-grained understanding of Chinese food culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19077–19095, Miami, Florida, USA, November 2024e. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1063. URL https://aclanthology.org/2024.emnlp-main.1063/.

3. Wenyan Li, Jiaang Li, Rita Ramos, Raphael Tang, and Desmond Elliott. Understanding retrieval robustness for retrieval-augmented image captioning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9285–9299, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.503. URL https://aclanthology.org/2024.acl-long.503/.

4. Wenyan Li, Raphael Tang, Chengzu Li, Clemente Pasti, Vésteinn Snæbjarnarson, Caiqi Zhang, Ivan Vulić, Ryan Cotterell, and Anders Søgaard. Lost in embeddings: Information loss in vision-language models. *Under review*, 2025b.

I also contributed to the following projects during my Ph.D., which are not included in this thesis:

1. Wenyan Li, Dong Li, Wanjing Li, Yuanjie Wang, Hai Jie, and Yiran Zhong. MAP: Low-data regime multimodal learning with adapter-based pre-training and prompting. In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 185–190, Gothenburg, Sweden, September 2023c. Association for Computational Linguistics. URL https://aclanthology.org/2023.clasp-1.19/.

2. Yong Cao, Wenyan Li, Jiaang Li, Yifei Yuan, Antonia Karamolegkou, and Daniel Hershcovich. Exploring visual culture awareness in gpt-4v: A comprehensive probing. *arXiv preprint arXiv:2402.06015*, 2024b. URL https://arxiv.org/abs/2402.06015.

3. Raphael Tang, Crystina Zhang, Lixinyu Xu, Yao Lu, Wenyan Li, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. Words worth a thousand pictures: Measuring and understanding perceptual variability in text-to-image generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5441–5454, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.311. URL https://aclanthology.org/2024.emnlp-main.311/. [**Outstanding Paper Award**]

4. Jiaang Li, Yifei Yuan, Wenyan Li, Mohammad Aliannejadi, Daniel Hershcovich, Anders Søgaard, Ivan Vulić, Wenxuan Zhang, Paul Pu Liang, Yang Deng, and Serge Belongie. Ravenea: A benchmark for multimodal retrieval-augmented visual culture understanding. *arXiv preprint arXiv:2505.14462*, 2025a. URL https://arxiv.org/abs/2505.14462.

5. Li Zhou, Lutong Yu, Dongchu Xie, Shaohuan Cheng, Wenyan Li, and Haizhou Li. Hanfu-bench: A multimodal benchmark on cross-temporal cultural understanding and transcreation. *arXiv preprint arXiv:2506.01565*, 2025. URL https://arxiv.org/abs/2506.01565.

6. Yuchen Huang, Zhiyuan Fan, Zhitao He, Sandeep Polisetty, Wenyan Li, and Yi R. Fung. Cultureclip: Empowering clip with cultural awareness through synthetic images and contextualized captions. *Conference on Language Modeling*, 2025.

Table of Contents

Chapter 1

Introduction

We have witnessed the great breakthrough in Artificial Intelligence (AI) in the past few years, including the trending intelligent chatbots such as ChatGPT[1] and realistic image and video generation (Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Singer et al., 2022; Podell et al., 2023). These groundbreaking innovations have fundamentally transformed how AI systems process information, interact with human, and solve complex problems across various domains. Since its emergence in the 1950s (Turing, 1950), AI has experienced remarkable evolution (LeCun et al., 2015; Jordan and Mitchell, 2015), particularly accelerated by breakthroughs in machine learning that leverage vast internet-based datasets, such as Common Crawl[2], Conceptual Captions (Sharma et al., 2018) and LAION (Schuhmann et al., 2022), etc. The field has evolved from single-modality applications toward more human-like perception and reasoning, which makes multimodal understanding a critical frontier in AI research (Baltrušaitis et al., 2018; Bommasani et al., 2021). These multimodal systems promise to revolutionize numerous applications, from accessibility tools that describe visual content to visually impaired users (Gurari et al., 2020), to creative platforms that generate images from textual descriptions (Ramesh et al., 2021), as well as cross-cultural communication that requires understanding of linguistic and visual cultural differences (Mogrovejo et al.; Nayak et al., 2024).

While existing multimodal models have achieved remarkable capabilities in tasks like image captioning (Li et al., 2022a,b) and visual question answering (VQA) (Antol et al., 2015; Goyal et al., 2017; Singh et al., 2019). their real-world robustness remains brittle. Raw web data contains significant noise, and current filtering techniques designed to reduce this noise frequently sacrifice important aspects of data diversity (Nguyen et al., 2023). Consequently, visual components in the state-of-the-art systems, including GPT-4V, fail at distinguishing visual differences that are apparent to human (Tong et al., 2024). These subtle data variations such as compositional, cultural context differences frequently

[1] https://chat.openai.com/.

[2] https://commoncrawl.org/.

trigger erroneous behaviors (Liu et al., 2021a; Thrush et al., 2022; Nayak et al., 2024). Such fragility reveals a fundamental challenge in multimodal learning: while we can scale models to achieve high benchmark scores (Wu et al., 2022), our understanding of how data properties propagate through complex architectures and learning process to shape model behavior remains incomplete. Developing advanced data curation techniques and understanding how data properties and corresponding feature representation propagate through model architectures to impact decisions is critical for developing trustworthy multimodal models (Bommasani et al., 2021; Dehghani et al., 2023). This thesis, aims to bridge this critical gap.

Human cognition is inherently multimodal, integrating visual, auditory, and tactile signals to form coherent perceptions (Stein and Meredith, 1993). Achieving comparable robustness in multimodal systems demands addressing two connected challenges: 1) how to deal with the intrinsic characteristic in data that these models encounter, at both training and inference stages, and 2) how to interpret the often "black-box" behavior of the models that process it.

Data quality, diversity, and representativeness profoundly influence the effectiveness of multimodal models (Paullada et al., 2021; Dodge et al., 2021). Recent research highlights the significance of carefully curated datasets in enhancing model performance, particularly in tasks such as image captioning (Atliha and Šešok, 2020; Nguyen et al., 2023). Despite this progress, systematically curating training data in a dynamic manner while maintaining the size of the dataset remains underexplored.

Cultural context introduces an additional layer of complexity, influencing how models interpret and respond to multimodal inputs regarding the underlying cultural aspects. For example, the symbol of a dragon often represents luck and fortune in Chinese culture while typically symbolizing danger and evil in Western countries. Large disparities exist between understanding culture-specific and common concepts (Nikandrou et al., 2024; Mogrovejo et al.), and adapting cultural specific concepts in image generation is proved to be challenging for state-of-the-art generative models (Khanuja et al., 2024). Although recent vision-language models (VLMs) have started incorporating cultural knowledge to account for these crucial dimensions (Liu et al., 2025; Li et al., 2025a), fine-grained understanding of regional cultural variations, especially concerning visually similar but conceptually distinct entities, remains a major and unresolved challenge (Li et al., 2024e).

Another critical gap lies in understanding the internal mechanisms of modern VLMs. These models exhibit unexplained sensitivities and visual blind spots—such as copying irrelevant tokens from additional provided contexts (Ramos et al., 2023c) or discarding critical visual features during modality fusion (Tong et al., 2024; Rahmanzadehgervi et al., 2024). Standard evaluation metrics often miss these systematic failures, which reveal fundamental limitations in how VLMs represent and process information. Diagnosing such behavioral pathologies and identifying why and where the information flow breaks down is essential for building more transparent and reliable multimodal systems.

In response, in this thesis we investigate these three fundamental questions in multimodal research:

1. How can we develop data curation strategies that effectively address variations and mismatches in multimodal data during training to dynamically improve model robustness and performance?

2. How should we evaluate multimodal understanding within cultural contexts to ensure fair and comprehensive assessment of their capabilities?

3. How to identify and characterize behavioral blind spots in VLMs, particularly in retrieval-augmented systems and modality fusion components, and how can we overcome them?

Our investigation systematically traces the flow of information through the multimodal pipeline, from data curation to culture-aware evaluation of VLMs and behavioral interpretation. We analyze how visual and textual contexts are encoded, how features are represented and fused across modalities, and how these processes impact the final generated output across three tasks: image captioning, retrieval-augmented generation, and visual question answering. By localizing where and why multimodal interactions break, this comprehensive analysis provides the foundational insights and practical tools necessary to engineer the next generation of more reliable, transparent, and ultimately more capable multimodal systems.

1.1 Contribution

Bridging Data-Centric Gaps in Training

Current models treat training data homogeneously, ignoring sample difficulty, retrieval noise, and cross-modal misalignment. To address this, in *Chapter 3* (Li et al., 2024d), inspired by curriculum learning(Bengio et al., 2009; Kumar et al., 2010), we introduce dynamic data curation techniques for training image captioning models. Our approach automatically updates the training dataset, modulating learning difficulty based on current model status. Through methods including either removing a sample, replacing the caption in a sample, or generating a new image from existing captions, we demonstrate that dynamic data curation efficiently improves image captioning performance on standard downstream captioning datasets without increasing the total size of the training datasets.

Benchmarking Fine-grained Cultural Understanding

To address the gap in evaluating VLMs within fine-grained cultural contexts, *Chapter 4* (Li et al., 2024e) introduces the *FoodieQA* benchmark. This human-curated, high-quality multimodal resource features original images not crawled from the internet, ensuring evaluation fairness by preventing potential data contamination in pretrained state-of-the-art models. FoodieQA examines regionally-specific cultural knowledge within the food domain through challenging multiple-choice VQA tasks spanning multi-image, single-image, and text-only settings. Our analysis uncovers distinctive failure patterns in vision-language models, particularly among open-weights multimodal models, highlighting their limitations in culturally-grounded reasoning and multi-image perception.

Identifying Behavioral Blind Spots

We first diagnose unfaithful behaviors in retrieval-augmented vision-language models. In *Chapter 5* (Li et al., 2024c), we diagnose critical limitations in SmallCap (Ramos et al., 2023c)—a retrieval-augmented captioner that enriches context using captions from visually similar images. While this approach significantly boosts captioning performance, particularly for out-of-domain examples, we notice a key vulnerability: retrieved content frequently misleads predictions, undermining model robustness. Through controlled experiments, we identify a copying behavior over *majority tokens* using attribution analysis and attention visualization. We then propose and validate context sampling as an effective mitigation strategy.

We then turn to the fundamental problem of modality fusion in multimodal models, diagnosing information loss at the architectural level. In *Chapter 6* (Li et al., 2025b), we investigate the information bottleneck within VLM connectors through a two-stage analysis. First, we employ a k-nearest neighbor overlap ratio to validate a critical hypothesis: that the simple projection modules used as connectors fundamentally alter the geometric structure of the original visual representations. Having established that structural integrity is not preserved, we then introduce a patch-level embedding reconstruction loss to diagnose what specific visual information is degraded. This novel metric allows us to quantify and localize the degradation in the original image, providing direct visualizations and a clear explanation for downstream failures in VQA and image captioning.

Chapter 2

Background

Multimodal learning is the process of learning from various data types, such as images, text, and speech (Ngiam et al., 2011). In this chapter, we provide a brief overview of the popular datasets and models used in multimodal learning and evaluation, especially the ones that are included and discussed in the following chapters.

2.1 Multimodal Representations and VLMs

In this dissertation, we focus on multimodal learning with visual and text inputs. More specifically, with the remarkable abilities of large language models, our studies mainly involve vision language models that enable visual language interactions through large language models as backbones.[1]

2.1.1 Visual Representations

Prior to the introduction of the Vision Transformer (Dosovitskiy et al., 2021), the field of computer vision was dominated by Convolution Neural Networks (CNNs) (Krizhevsky et al., 2012), with progress driven by innovations in both architecture and training paradigms. The architectural advances was represented by models like EfficientNet (Tan and Le, 2019), which proposed a scaling method to optimize network depth, width, and resolution, setting a new state-of-the-art on image classification benchmarks. Concurrently, the performance of these advanced architectures was significantly boosted by self-supervised learning methods that learned rich features from unlabeled data. Prominent among these were contrastive learning frameworks such as MoCo (He et al., 2020) and SimCLR (Chen et al., 2020).

[1] In this dissertation, we do not consider VQ-VAE (Van Den Oord et al., 2017) based VLMs, which are more often used for text-to-image generation.

The introduction of the **Vision Transformer** (ViT) by Dosovitskiy et al. (2021) marked a pivotal moment in visual representation learning, establishing that a pure Transformer architecture (Vaswani et al., 2017) could match the performance of state-of-the-art CNNs. The core idea was to treat an image as a sequence of tokens by splitting it into fixed-size patches, which are then linearly embedded and processed by a standard Transformer encoder.

2.1.2 Large Language Models

The sophisticated capabilities of modern Vision-Language Models are built upon a series of foundational breakthroughs in Large Language Models (LLMs), which provide their reasoning and linguistic core. The paradigm shift began with the Transformer architecture (Vaswani et al., 2017). By replacing the sequential processing of Recurrent Neural Networks (Hochreiter and Schmidhuber, 1997; Chung et al., 2014) with a highly parallelizable self-attention mechanism, the Transformer allowed models to process entire sequences at once, capturing complex dependencies between words regardless of their distance. This architectural innovation unlocked the ability to train on web-scale data. This led to the era of large-scale generative pre-training, exemplified by models like GPT-3 (Brown et al., 2020) and T5 (Raffel et al., 2020). Trained on vast internet corpora with a simple objective like next-token prediction, these models became powerful parametric knowledge bases, storing factual information, common sense, and linguistic patterns within their weights. Critically, this massive pre-training gave rise to an emergent capability: in-context learning. This allowed the models to perform a wide array of new tasks simply by being prompted with a few examples in natural language, without requiring any updates to their weights. This discovery was critical for early VLMs like Frozen (Tsimpoukelli et al., 2021), as it established that a static, pre-trained LLM could learn to interpret novel inputs, such as digesting projected visual embeddings as part of its contexts.

Subsequent progress focused on optimizing and aligning these powerful models. Improved scaling laws (Hoffmann et al., 2022) provided a more efficient recipe for balancing model size with training data volume, leading to better performance for a given compute budget. Then, pioneered by FLAN (Wei et al., 2022) and later refined in models like InstructGPT, large-scale instruction tuning was introduced. Instruction tuning involved fine-tuning the base LLM on a massive, diverse collection of tasks formatted as explicit instructions (e.g., questions, summaries, translations). This process transformed the models from simple text-completion engines into flexible and controllable instruction-following agents. Finally, by aligning with user intent, such instruction-following LLMs provide a strong general-purpose reasoning backbone that could be directly integrated into more complicated multimodal models like BLIP-2 (Li et al., 2023b) and LLaVA (Liu et al., 2023a), allowing them to complete advanced conversational and instruction-following tasks based on both text and image inputs.

2.1.3 Multimodal Representations

Building modern vision-language models involves solving two key challenges: first, converting an image into a representation that is semantically compatible with language, and second, interfacing this representation with a powerful LLM.

The first challenge was pivotally addressed by Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021). Rather than just extracting generic visual features, CLIP's vision encoder is trained alongside a text encoder on 400 million image-text pairs. This contrastive process forces the model to learn a shared embedding space where visual concepts are explicitly aligned with their linguistic descriptions, producing image embeddings that are inherently language-aware.

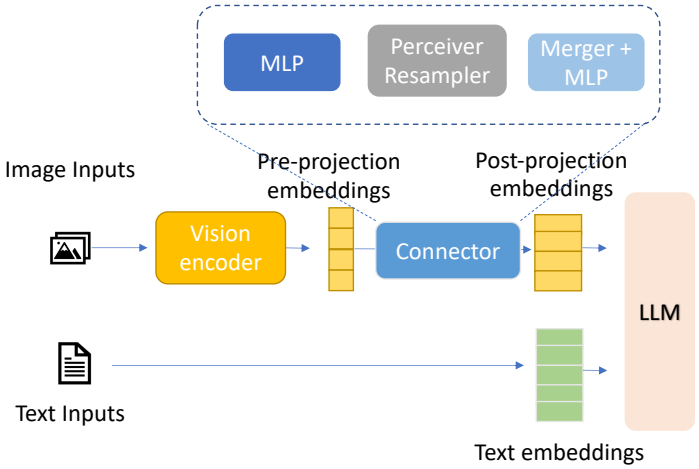


Figure 2.1: Connector-based modality fusion for vision language models.

Addressing the second challenge: how to feed rich visual features in a compatible format that an LLM could digest, led to various vision-language fusion strategies, broadly categorized into shallow and deep approaches (Li and Tang, 2024). Shallow fusion, known for its efficiency, involves using a simple connector module—typically a Multi-Layer Perceptron (MLP)—to project visual embeddings into the LLM's input space. Figure 2.1 illustrates a general structure of connector-based vision-language models. After projection, the input image could be used as a static prefix to the text prompt. The pioneering work by Tsimpoukelli et al. (2021) first demonstrated this approach with the Frozen model, showing that by training only a small mapping network, a frozen LLM could learn to digest images while keeping its vast knowledge intact. In contrast, more sophisticated strategies were developed for a more intricate integration. For example, Flamingo (Alayrac et al., 2022) employs deep fusion through cross-attention mechanisms that allow visual information

to influence each layer of the LLM. Others, like BLIP-2 (Li et al., 2023b), use advanced adapter modules, such as the Q-Former, to allow visual features to dynamically interact with the LLM's hidden states throughout its processing. By leveraging an instruction-tuned LLMs such as FLAN-T5 (Chung et al., 2024) or OPT (Zhang et al., 2022b), BLIP-2 enables the vision-language model to complete instruction-following tasks. With the LLM's pre-existing reasoning and generative abilities, these VLMs achieve strong zero-shot capability and can perform complex multimodal tasks like detailed image description, visual question answering, and multimodal dialogue.

The move towards instruction-following capabilities was further significantly advanced by LLaVA (Liu et al., 2023a), which pioneered a shift beyond the strictly frozen paradigm. LLaVA introduced end-to-end training, updating the weights of both modality projection layers and also the pretrained LLM on a dataset of multimodal instructions. This methodology sacrificing some frozen efficiency for significantly improved conversational ability has since become a dominant approach. It has led to a new wave of powerful open-source models like Qwen-VL (Bai et al., 2023), known for its strong performance, and Idefics models (Laurençon et al., 2024), which extends the instruction-following capability to handle arbitrarily interleaved sequences of images and text.

2.1.4 Retrieval Augmentation in VLMs

A primary strategy for improving Vision-Language Models (VLMs) is to scale their internal, parametric knowledge by training ever-larger models on larger datasets (Wang et al., 2025). In contrast, an alternative and complementary approach has emerged with Retrieval-Augmented Vision-Language Models (RA-VLMs). By combining internal knowledge with external, non-parametric information, RA-VLMs function as "open-book" systems that can consult verifiable evidence before generating a response.

The application of this paradigm has evolved from specialized, lightweight models to general-purpose reasoning engines. For instance, SmallCap (Ramos et al., 2023c) demonstrated a highly parameter-efficient approach to image captioning. Instead of end-to-end training of a large VLM, it uses retrieved captions from visually similar images to form a textual prompt for a frozen, off-the-shelf language model, guiding it to generate a high-quality description. Following this, the approach was scaled to tackle open-domain, knowledge-intensive tasks. REVEAL (Hu et al., 2023) showed that retrieving multimodal documents could enable VLMs to answer complex questions requiring external knowledge. More advanced architectures like RAC-M3 (Yasunaga et al., 2023) have further leveraged web-scale retrieval for generating both texts and images.

Challenges

Despite remarkable advancements in architecture and performance, interpretability remains a fundamental challenge in multimodal systems. At the modality fusion stage, components like MLPs or Q-Formers operate as black boxes, making their internal processes opaque. The transformation of detailed pixel-level data into abstract language embeddings represents an irreversible and lossy compression process. This opacity creates several difficulties: verifying preservation of critical visual information, understanding error sources, and tracing reasoning back to specific image regions becomes nearly impossible.

Similarly, when models employ retrieval augmentation, the influence of retrieved content on decision-making remains unclear—whether it aids or misleads the final output is difficult to determine. In both modality fusion and retrieval scenarios, the model's reasoning pathways are obscured, potentially resulting in plausible but visually ungrounded predictions (Rahmanzadehgervi et al., 2024). This lack of transparency constitutes a significant barrier to developing robust and reliable vision-language models, a challenge this thesis aims to address.

2.2 Benchmarking VLMs

Comprehensive evaluation of VLMs requires multifaceted benchmarks that assess different capabilities across diverse scenarios. This section outlines the tasks and datasets used for evaluating and benchmarking multimodal models, with a specific emphasis on those involved in this thesis. Early research focused on foundational tasks such as image captioning, where models learn to generate a descriptive sentence for an image, and Visual Question Answering (VQA), where they must answer a natural language question about visual content. These capabilities were primarily developed and evaluated on curated, human-annotated datasets.

2.2.1 Image Captioning

Widely used image captioning dataset include MS-COCO (Chen et al., 2015) and Flickr30k (Young et al., 2014), which provided clean, high-quality image-text pairs. For evaluating out-of-domain performance, datasets like VizWiz (Gurari et al., 2020) and NoCaps (Agrawal et al., 2019) are frequently used. VizWiz centers on images captured by visually impaired users, requiring captions that are practical and useful while often involves challenges like blurry images or unanswerable questions; NoCaps explicitly tests a model's ability to generalize by captioning novel objects that do not appear in the MS-COCO training set. Table 2.1 summarizes the use case and size of these captioning datasets.

Captioning Dataset	Images	Domain	Training	Evaluation
MS-COCO (Lin et al., 2014)	123k	In-Domain	✓	✓
Flickr30K (Young et al., 2014)	31k	In-Domain	✓	✓
VizWiz (Gurari et al., 2020)	39k	Out-of-Domain		✓
NoCaps (Agrawal et al., 2019)	15k	Out-of-Domain		✓

Table 2.1: Image Captioning Datasets.

2.2.2 Visual Question Answering

The evaluation of Visual Question Answering (VQA) capabilities has evolved from foundational benchmarks to more comprehensive and grounded assessments. VQAv2 (Goyal et al., 2017) serves as the primary in-domain benchmark for testing a model's ability to answer questions about an image's content (e.g., "What color is the car?"). While crucial for establishing baseline performance, it primarily assesses recognition and simple reasoning. To measure the more sophisticated capabilities of modern instruction-tuned VLMs, comprehensive benchmark suites like SEED-Bench (Li et al., 2024b) have been introduced. SEED-Bench moves beyond simple VQA, evaluating models across a diverse range of tasks, including counting, spatial understanding, complex reasoning, etc. Targeting at real-world applicability, the VizWiz-Grounding benchmark (Chen et al., 2022) was developed. Sourced from images taken by visually impaired users, it requires a model to not only answer a question but also to ground its answer by providing bounding box coordinates for the relevant objects. This directly evaluates whether a model's response is verifiably linked to the visual evidence, making it valuable for examining the problem of hallucination and measuring practical reliability.

2.2.3 Culture-related Benchmarks

Recently, there has been growing interest in evaluating cultural awareness and bias in VLMs, advancing from foundational perception tasks to more complex, contextually grounded reasoning. The challenge begins at the level of object recognition, as highlighted by benchmark like GeoDE (Ramaswamy et al., 2023), which demonstrates that standard vision encoders struggle to recognize objects from non-Western geographies. Moving beyond recognition, MaRVL (Liu et al., 2021a) employs a minimal-pair design to test a model's ability to understand relationships between culturally specific concepts, requiring genuine visio-linguistic reasoning rather than simple factual recall.

Building on this foundation, question-answering datasets have been developed to directly assess explicit cultural knowledge. For example, CVQA (Romero et al., 2024) and the larger-scale CultureVQA (Nayak et al., 2024) focus on visual question answering

tasks centered on culturally specific content. Additionally, domain-specific benchmarks like WorldCuisines (Winata et al., 2025) evaluate cultural awareness within specialized areas, such as food recognition at a global level.

2.2.4 Other Datasets

Beyond what has been discussed above, a diverse suite of benchmarks has been developed to test more advanced VLM capabilities. Foundational benchmarks like Caltech-UCSD Birds-200-2011 (CUB) (Wah et al., 2011) probes fine-grained understanding between visually similar subcategories, such as distinguishing one species of sparrow from another. To probe deeper reasoning, benchmarks like SNLI-VE (Xie et al., 2019) and Winoground (Thrush et al., 2022) evaluate logical and compositional understanding, while grounding tasks like RefCOCO (Kazemzadeh et al., 2014) verify if a model can link language to specific visual evidence. Meanwhile, the emergence of instruction-tuned VLMs has led to comprehensive evaluation suites such as MME (Fu et al., 2023) and SEED-Bench-2 (Li et al., 2024a), designed to assess reasoning abilities across diverse tasks.

With massively available data online, training datasets also have shifted from task-specific datasets to large-scale, web-crawled image-text pairs. Datasets like Conceptual Captions (Sharma et al., 2018) and the LAION (Schuhmann et al., 2022, 2021) series are commonly used for pre-training, enabling models like CLIP (Radford et al., 2021) and BLIP-2 (Li et al., 2023b) to learn a robust vision-language alignment from billions of examples. While this web-scale data is inherently noisy, its large scale provides the foundation for general visual and language understanding. More recent advancement that builds upon this foundation includes instruction-tuning datasets like LLaVA-Instruct-150k.[2] These synthetically generated collections of multimodal dialogues teach a pre-trained VLM to follow commands and engage in complex reasoning. However, as both web-crawled and synthetic data introduce significant noise, data curation techniques are critical to modern VLM training pipelines.

Challenges

Despite recent progress, critical gaps remain in VLM benchmarking, particularly in two directions. First, fine-grained cultural understanding remains largely unexplored. While models can differentiate biological categories, they struggle with culturally significant variations, such as reasoning about the taste and origin of regional food. Second, a related frontier is multi-image comparative reasoning, where models are tasked with multi-hop reasoning that requires differentiating between multiple, similar images based on specific fine-grained attributes. Both challenges highlight a need to move beyond single-image

[2] https://huggingface.co/datasets/liuhaotian/LLaVA-Instruct-150K

recognition toward a deeper, comparative, and contextualized form of visual reasoning that current benchmarks are only beginning to address.

Chapter 3

The Role of Data Curation in Image Captioning

Abstract

Image captioning models are typically trained by treating all samples equally, neglecting to account for mismatched or otherwise difficult data points. In contrast, recent work has shown the effectiveness of training models by scheduling the data using curriculum learning strategies. This paper contributes to this direction by actively curating difficult samples in datasets without increasing the total number of samples. We explore the effect of using three data curation methods within the training process: complete removal of an sample, caption replacement, or image replacement via a text-to-image generation model. Experiments on the Flickr30K and COCO datasets with the BLIP and BEiT-3 models demonstrate that these curation methods do indeed yield improved image captioning models, underscoring their efficacy.

3.1 Introduction

Image captioning is the task of generating grammatically correct and accurate descriptions of visual data, which involves understanding the identity of salient objects and their relationships (Bernardi et al., 2016; Baltrušaitis et al., 2018). While existing models have made significant progress on this problem, there remains an inherent challenge: how to address the variations in learning difficulty that arise from diverse image-caption pairs (Sharma et al., 2018; Schuhmann et al., 2021).

Image captioning models are usually trained by treating the entire training dataset equally, which overlooks the variations in the complexity of each data point. One attempt at addressing this issue has been to apply data filtering as a preprocessing stage to large-

scale datasets to remove noisy data from the pretraining process (Li et al., 2022b; Nguyen et al., 2023). Several other image captioning techniques have relied on curriculum learning strategies (Bengio et al., 2009), which schedule the training data with increased levels of complexity, effectively adapting the learning process to the difficulty of the task (Liu et al., 2021b; Dong et al., 2021; Zhang et al., 2022a; Alsharid et al., 2021; Ayyubi et al., 2023). In this paper, we aim to answer a fundamental question: **can image captioning models be improved by not only recognizing variations in the data but also actively curating difficult samples?**

We introduce three data curation methods, each with the aim of improving the learning process while preserving the overall size of the training dataset. These methods include the complete removal of a sample, the replacement of captions, or the substitution of images using a text-to-image generation model. The targets of these methods are image-caption training samples that have unusually high losses with respect to the rest of the training dataset under the current model parameters. In other words, our approach focuses on the samples that are proving *difficult* to model (Bengio et al., 2009; Kumar et al., 2010).

The main findings of this paper are:

- Dynamic data curation enhances image captioning performance. The best strategy varies between datasets but is generalizable to different vision-language models.[1]
- The extent of curation is a critical factor and dataset dependent. We find that curating more than 50% of data negatively impacts the effectiveness of data curation.
- Image generation-based curation has potential benefits with specific techniques, but its potential benefit is limited by generation errors identified through a human study, which are not apparent from automatic evaluation metrics, such as CLIPScore (Hessel et al., 2021).

3.2 Related work

Data Curation in NLP While still under-explored for image captioning, Rogers (2021) highlighted the importance of data curation for deep learning and NLP. Several studies have adopted data curation for large language models: Chen et al. (2023) developed a general text curation framework based on large language models; Kandpal et al. (2022) and Lee et al. (2022) discussed the impact of deduplication for training; Chang and Jia (2023) shows that careful curation alone can stabilize in-context learning.

Image Captioning and Learning Strategies Curriculum learning (Bengio et al., 2009) and self-paced learning (Kumar et al., 2010) are techniques that adjust the learning

[1]We release the code for our curation framework at https://github.com/lyan62/data-curation/

process based on variations in the learning samples, leveraging loss values to estimate model competence. For image captioning, several studies have introduced diverse learning techniques aimed at customizing the model training process in terms of sample difficulty, incorporating both textual and visual features (Alsharid et al., 2021; Dong et al., 2021; Zhang et al., 2022a). Whereas these methods adjust model training using sorted data, our approach proposes an innovative perspective: adjusting training by curating data samples that exhibit outlier losses, while preserving the overall dataset size.

Text-to-image Generative Models Text-to-image generative models, including diffusion models (Song et al., 2021; Nichol and Dhariwal, 2021), have rapidly gained popularity and proven powerful. Although recent large-scale latent diffusion models excel in generating high-resolution images with artistic and photo-realistic qualities (Rombach et al., 2022; Nichol et al., 2022; Ramesh et al., 2022; Saharia et al., 2022), their application in multi-modal tasks remains unexplored. Concurrently to our work, Azizi et al. (2023) and Jain et al. (2023) show that image classifiers can be improved by learning from augmented images generated by finetuned generative models; Xiao et al. (2023) and Caffagni et al. (2023) used generative models to augment the datasets used to train captioning models.

To the best of our knowledge, we are the first to explore how dynamic data curation approaches can impact downstream image captioning *without* scaling up existing datasets, and how text-to-image generative models can be applied in the process.

3.3 Data Curation for Captioning

Our main goal is to assess whether actively curating image-caption pairs during training can improve image captioning models. There are many reasons for the existence of difficult samples, including mismatches between the image-caption or inconsistencies between the image and caption (Atliha and Šešok, 2020), e.g. the caption includes mentions of entities that cannot be seen in the image. For clarity in what follows, let $\mathcal{D}$ be an image captioning training dataset with K images, and let I_k be the k-th image. Each image is paired with J captions; let C_k^j be jth caption of image k, and thus, let $(\mathrm{I}_k, \mathrm{C}_k^j)$ be an image–caption sample.

3.3.1 Identifying the difficult samples

Inspired by scheduling in curriculum learning (Bengio et al., 2009; Kumar et al., 2010), we assume that difficult training samples can be automatically identified throughout the training process. We propose to use the captioning model $\mathcal{M}$ that is being trained on dataset $\mathcal{D}$ to automatically identify such samples. We can readily use this model to calculate the loss of each sample in $\mathcal{D}$ at any point in time, such as at the end of each epoch

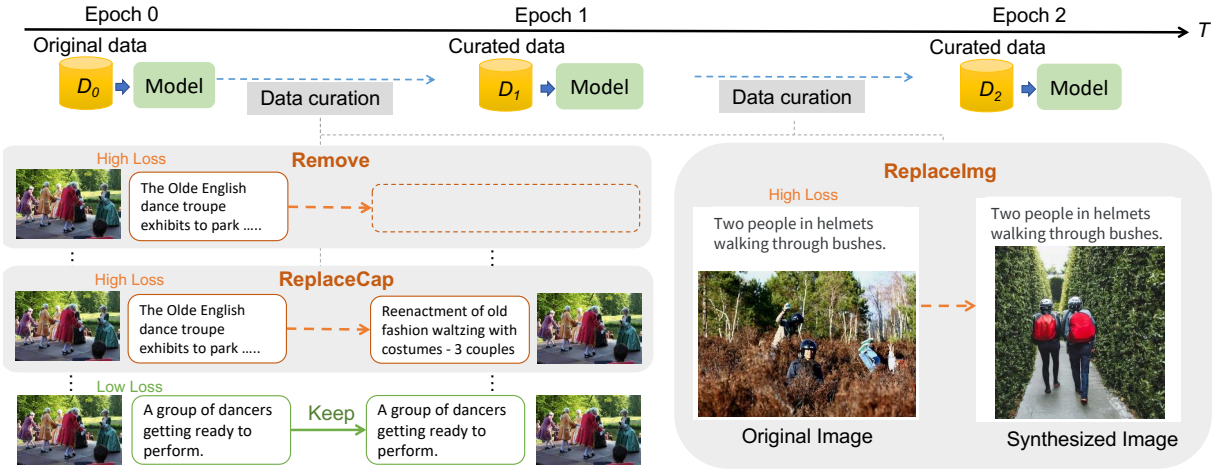


Figure 3.1: Overview of our data curation methods. For REMOVE, high loss image-text pairs are removed; for REPLACECAP, the image is paired with an alternative caption from the original dataset; for REPLACEIMG, captions of original images are used as prompts for text-to-image generation to synthesize new image–text pairs. We experiment with both options of replacing the image only, or pair another relevant caption to the synthesized image.

t: $\mathcal{L}^t_{\mathcal{M}}(\mathrm{I}_k, \mathrm{C}^j_k)\ \forall j, k$. The samples can be be ranked by their respective losses, providing candidates for samples that may benefit from data curation. In particular, the highest loss samples are targets for our data curation methods. We focus on samples with losses that are either two standard deviations from the mean, or the top $X\%$ highest loss samples. The data curation performs dynamic updates to the training dataset $\mathcal{D} \to \mathcal{D}_1 \to \cdots \to \mathcal{D}_T$. In this way, the training dataset is dynamically updated at the end of each epoch according to the model's current captioning capability at time t. We empirically observe that without data curation, the high-loss samples remain high-loss during five epochs of training.[2]

3.3.2 Curation approaches

We investigate three approaches to dynamically curate the high-loss image-caption pairs: REMOVAL, REPLACECAP, and REPLACEIMG. Figure 3.1 shows an overview of these approaches.

Remove The simplest approach to data curation is to remove the high-loss samples, preventing the samples from confusing the model. In REMOVE, the high-loss samples are

[2]The leftmost plot in Figure 3.2 shows the empirical distribution of losses in the training samples of the Flickr30K.

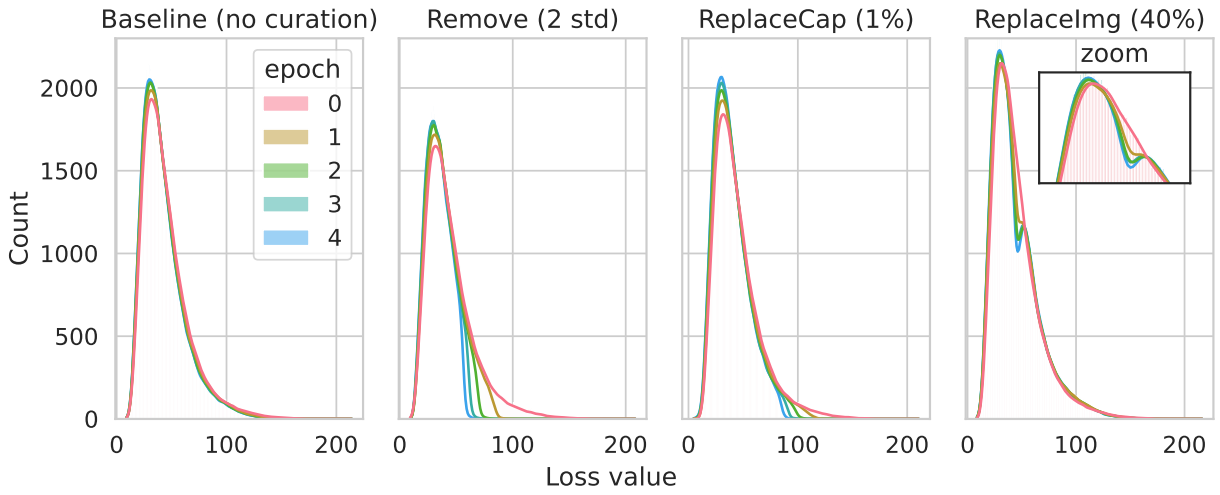


Figure 3.2: Different curation methods change the loss distribution of training samples over epochs for Flickr30K. In contrast, in the absence of data curation (the leftmost plot), high-loss samples consistently retain their high-loss status throughout training.

completely removed from the remainder of the training process, reducing the total number of image–caption training samples.

ReplaceCap In ReplaceCap, we simply replace the caption in the image–caption sample with a different caption from the original dataset that describes the image, effectively creating a duplicate. With this method, the total number of samples used to train the model remains the same, as well as the total number of the unique images. This creates a control condition for our experiments. As an alternative, we also experiment with replacing the original caption with one generated by a language model, which we discussed in Section 3.6.

ReplaceImg In ReplaceImg, we perform data curation using a text-to-image generative model. This has the benefit of training the model on the same total number of samples while exposing it to more unique images. In a rapid model-in-the-loop step, we use a text-to-image generation model to synthesize images based on the other sentences that describe the image. We integrate this into training as follows: Given an image I_k in the training data and its captions $\{(I_k, C_k^1), \ldots, (I_k, C_k^J)\}$, we synthesize a new image $\hat{I}_k$ without increasing the total number of samples in the original dataset. Specifically, for image I_k, we replace an original high-loss sample (I_k, C_k^j) with the synthesized image-text pair $(\hat{I}_k, C_k^j)$.

Given a set of captions that describe an image, there are several options for how to prompt the image generation model (Figure 3.11 in Appendix). We experiment with three

options:

- Single caption: Each caption is used in isolation to generate a new image.
- Sentence-BERT selection: There is a lot of variety in how different captions describe the same image. Instead of using all captions, we can use a representative caption from the set. This is achieved using the Sentence-BERT (Reimers and Gurevych, 2019) model to find the caption that is closest to the average embedding of all captions.
- Concatenation: All five captions are concatenated as the text prompt for generation.

For all three approaches mentioned above, we can append an additional string to the prompt as a *styler* to force a specific style in the generated image (+Styler). The styler used here is: "national geographic, high quality photography, Canon EOS R3, Flickr".[3] Some representative examples of images generated using this technique can be seen in Figure 3.13 in the Appendix.

3.4 Experimental Setup

3.4.1 Data & Metrics

We evaluate our data curation methods during finetuning on the widely used MS COCO (Lin et al., 2014) and Flickr30K (Young et al., 2014) datasets. We report results using the metrics of **B**LEU (Papineni et al., 2002), **M**ETEOR (Denkowski and Lavie, 2014), **C**IDEr (Vedantam et al., 2015), and **C**LIP**S**core (Hessel et al., 2021).

3.4.2 Models & Implementation

Image Captioning Models We study the effectiveness of data curation with two state-of-the-art pretrained vision-language models – BLIP (Li et al., 2022b) and BEiT-3 (Wang et al., 2023a). We note that BLIP has a captioning and filtering (CapFilt) data augmentation process during its pretraining, where both components were finetuned on the COCO dataset. Therefore we use pretrained checkpoint $\text{BLIP}_{CapFilt}$ for Flickr30k and BLIP_{base} for COCO in our experiment, removing the effects of the CapFilt process. We finetune BLIP using a total batch size of 128 for 5 epochs on 4×A100 GPUs. The BEiT-3 base model is finetuned with the default setups: a total batch size of 256 for 10 epochs on 8×A100 GPUs.

[3]The styler was chosen by inspecting the generated images, with a preference for photographic outputs and against "artistic" outputs, such as sketches and computer art.

			BLIP				BEiT-3			
	Method	Ratio	B	M	C	CS	B	M	C	CS
Flickr30K	Baseline	-	37.6	27.2	92.8	78.6	28.9	27.2	79.3	**80.4**
	+Remove	2 std	38.6	**27.4**	**95.8**	**79.2**	31.4	27.1	**83.7**	80.0
	+ReplaceCap	1%	37.9	**27.4**	94.5	78.9	29.6	**27.5**	80.1	80.3
	+ReplaceImg	40%	**39.0**	27.3	95.7	79.1	**32.0**	26.9	82.4	79.1
COCO	Baseline	-	39.9	30.8	132.0	77.3	39.4	31.1	133.7	77.4
	+Remove	1%	40.1	30.9	132.5	77.3	39.3	31.1	133.2	77.3
	+ReplaceCap	1%	**40.2**	30.9	132.7	77.3	39.4	31.0	133.6	76.5
	+ReplaceImg	10%	**40.2**	**31.0**	**133.1**	77.3	**39.6**	**31.1**	**134.4**	**77.5**

Table 3.1: Results of finetuning with our data curation methods compared to standard finetuning of BLIP and BEiT-3 on the Flickr30K and COCO datasets. We report **B**LEU, **M**eteor, **C**IDEr, and **C**LIP**S**core. Best scores are in **bold**.

Curation Ratio We tune the amount of data to be curated for each method on the validation data of each dataset using the BLIP model. See Section 3.6 for more discussion on the trade-off between the amount of data curation and model performance.

ReplaceImg Text-to-image Generation For text-to-image generation in ReplaceImg, we use the open source Stable Diffusion model (Rombach et al., 2022), which can generate images given a textual prompt. We finetune a Stable Diffusion v1.5 model, using the MS COCO (Lin et al., 2014) dataset with a prompt consisting of a concatenation of all 5 captions, for 15,000 steps with a constant learning rate of $1e{-}5$ and a batch size of 32. We experiment different versions of the released Stable Diffusion models and various techniques for generating high-quality images for replacement.[4] We find that using a finetuned text-to-image model enhances image captioning performance. See Section 6.9.3 for further analysis and ablation.

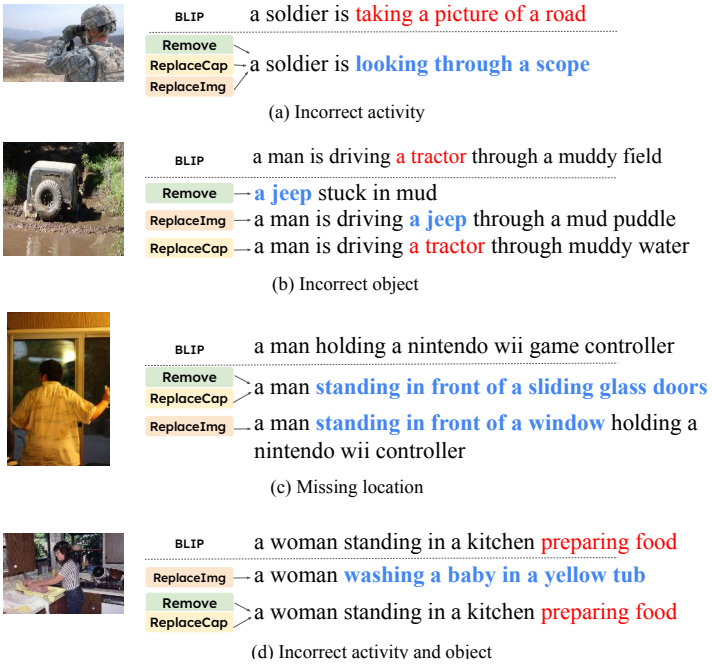


Figure 3.3: Qualitative examples from the COCO dataset of captions generated by the BLIP model (top), and the same models trained using our data curation methods (bottom). After curation, many of the errors (in red) can be avoided or fixed (in blue).

3.5 Results

3.5.1 Data curation improves captioning

Table 3.1 shows the results for the Flickr30K and COCO datasets with the BLIP and BEiT-3 models. The main conclusion is that better model performance can almost always be achieved using data curation. For Flickr30K, it can be seen that Remove (2 std) and ReplaceImg (40%) perform similarly well with a 2.9–3 CIDEr points improvement. The ReplaceCap method only improves performance by 1.7 CIDEr points when applied to the top 1% of high-loss samples. For COCO, the best performing approach is ReplaceImg with a curation ratio of 10%, bringing a 1.1 CIDEr point improvement over the baseline. ReplaceCap and Remove both work best when curating the top 1% of high-loss samples, bringing smaller improvements of 0.5–0.7 CIDEr points. Qualitative examples of the improvements can be seen in Figure 3.3.

[4]It is also possible to use API-based models but we chose Stable Diffusion because (i) Stable Diffusion can be integrated directly into our training pipeline using the open source code. And (ii) we estimate that it would cost $4,176 to run a single experiment on the Flickr30K dataset using DALL·E-2 as of Feburary 1st, 2024.

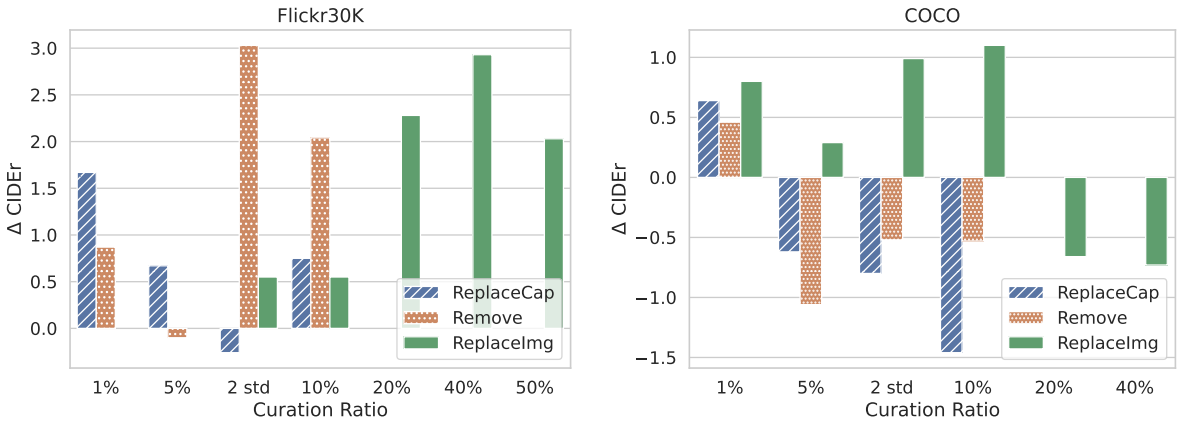


Figure 3.4: Effects of varying the amount of data curated. We observe that Flickr30K needs more curation (40% ReplaceImg or 2 std Remove) than COCO (10% ReplaceImg or 1% ReplaceCap). Flickr30K benefits more from removing high-loss training samples, indicating the original dataset may be noisier than MS COCO. For the 2 std approach, the number of samples curated is not fixed after each epoch and varies between 5% to 10%.

3.5.2 Generalization to different VL models

We also verify that our data curation methods generalize to other models by implementing them in the BEiT-3 model. More specifically, we used exactly the same curation ratio that gained improvements for BLIP. As shown in Table 3.1, where Remove is also the most efficient approach for better captioning on Flickr30K, and ReplaceImg improves the most for COCO. This shows that the curation methods can be readily applied to other state-of-the-art vision-language models and the curation ratios are transferable. We note that since BEiT-3 includes COCO in pretraining, the Remove and ReplaceCap methods are not beneficial.

3.6 Discussion

Curation amount matters The amount of data curated is an important hyperparameter. In addition to the best results reported above, we present finer-grained results of varying the amount of data curation. For Remove and ReplaceCap, we explore curating the top 1%, 5% and 10% of high-loss samples. For ReplaceImg, we explore 10%–80% curation ratios. In addition to fixed X% ratios, we also intervene on samples that have losses two standard deviations worse than the mean.

The results of this analysis are shown in Figure 3.4. While the effective curation ratio for different curation approach ranges from 1%-50% for Flickr30K, COCO benefits from

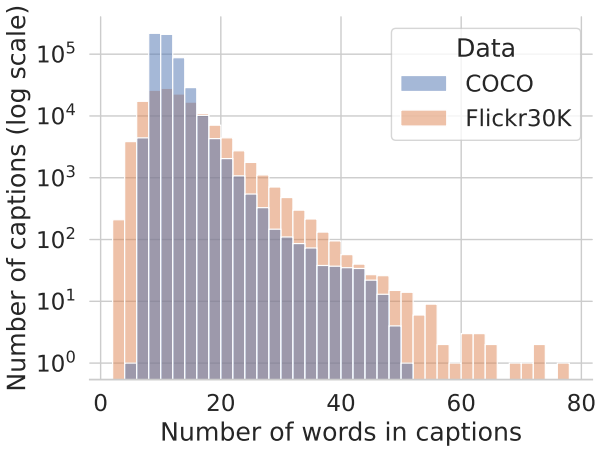


Figure 3.5: Distribution of caption lengths.

ReplaceImg on less than 10% of the top loss samples, and the effective curation ratio for Remove and ReplaceCap stops at 1%. This indicates that Flickr30K may contain more noisy samples than the MS COCO dataset. Compared to MS COCO, Flickr30K contains more samples with long captions (Figure 3.5), which may include overly-specific details that are inconsistent with other captions and are hard for the model to learn (Figure 3.12). Through our curation-based finetuning, these samples can be effectively identified, removed or replaced, which indicates that our method is efficient when training with noisy datasets. We note that curating more than 50% of the data does not benefit training and actually harms performance.

Curation changes training distributions We examine the loss distributions of training samples across epochs for each curation method to understand their impact on the training process (Figure 3.2). These losses are computed after each epoch using the current model parameters, with high-loss samples being targeted for the subsequent curation step.

For the Remove approach, training samples with loss that are two standard deviations worse than the mean are dynamically removed during training, leading to the shrinking tail of the loss distribution. ReplaceImg gradually reduces losses, resulting in the losses forming a mixture of Gaussians consisting of the original image-text pairs and the those with synthesized images. Going beyond just the losses of the training samples, we also inspect the distributions of the words in the training captions for the curation methods. Figure 3.6 shows these distributions, where it can be seen that Remove reduces low-frequency and singleton words during training, while ReplaceCap increases the counts of some lower-frequency words while removing singletons. By definition, ReplaceImg only changes the distribution of the images used to train the model, and as such, does not change the distribution of the words in the training data.

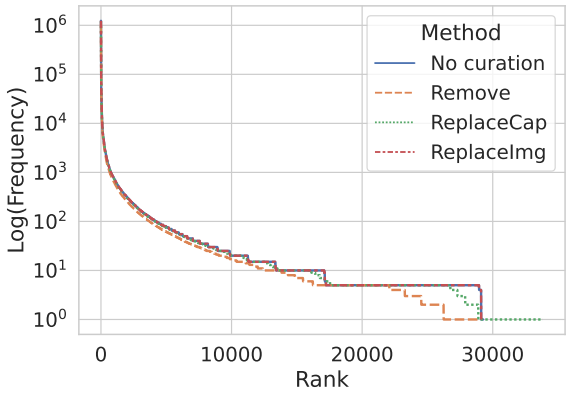


Figure 3.6: Zipfian distribution of words in Flickr30K training samples for different curation approaches. Note the clear changes made to the tail by REMOVE.

The efficacy of dynamic replacement Using training loss values as an effective indicator, we dynamically curate on the training samples identified as challenging. In REPLACEIMG, another static approach is to replace the identical images, i.e. I_k in $\{(I_k, C_k^1), \ldots, (I_k, C_k^J)\}$, with unique synthesized images before training, instead of updating the training samples while training. With static image replacement, for each of the reference captions, we replace their original image with a generated image. Static replacement with 20%–80% curation ratio corresponds to replacing images for one–four captions of the original five. The 50% replacement ratio mimics a fair coin-flip, where for each of the text-image samples, there is 50% probability for the image to be replaced by a synthesized image.

We compare the efficacy of these two approaches in Figure 3.7. When evaluating on the original 1k validation set, we see that for both approaches, incorporating synthesized images of 20% or 40% can assist finetuning and achieves higher CIDEr scores. Nevertheless, dynamic image replacement consistently performs better than the static method, showing focusing on the hard samples is effective. For both replacement methods, performance starts to decrease when the curation ratio is too high. This may indicate that when incorporating too many images from the synthetic distribution, the gap increases between the training and evaluation sets.

Replacing captions with LM generations As an alternative to the REPLACECAP method, we investigate the utility of replacing the captions with those generated by a language model (LM). Inspired by the approach in Ramos et al. (2023b), we prompt the XGLM-2.9B model (Lin et al., 2022) with few-shot examples to generate a new caption. The LM generated caption is then paired with the image as the curated sample. We

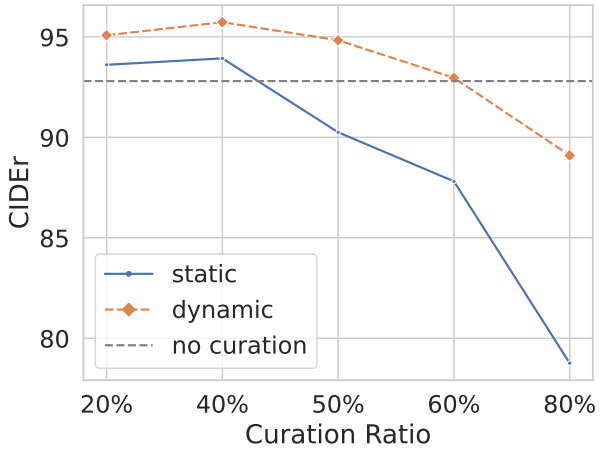


Figure 3.7: Dynamic versus static replacement for ReplaceImg using BLIP on the Flickr30K dataset, as a function of the number of samples replaced.

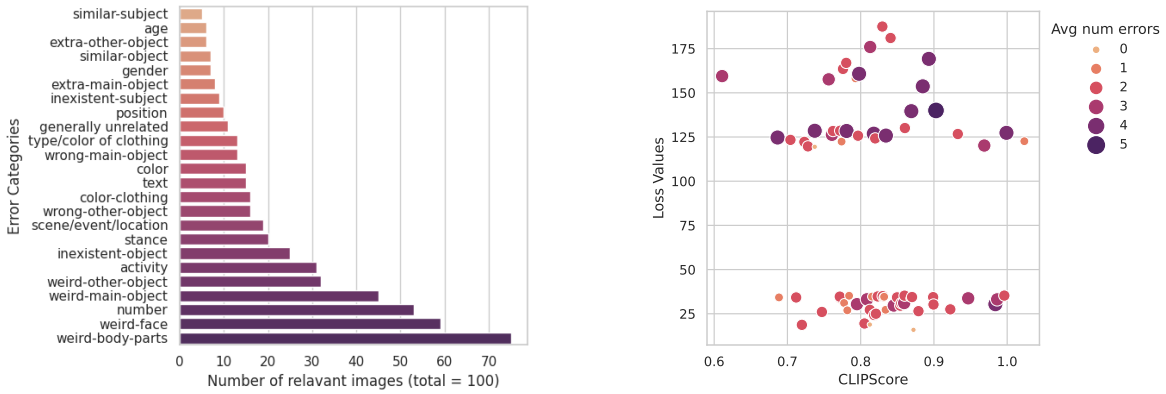
evaluate on Flickr30K using both models, applying the same curation ratio of 1% as ReplaceCap. The results presented in Table 3.2 indicate that this approach can serve as a viable alternative to ReplaceCap, consistently outperforming baselines for both models. Please refer to Appendix 3.9.3 for more implementation details.

	BLIP		BEiT3	
Method	BLEU	CIDEr	BLEU	CIDEr
Baseline	37.6	92.8	29.8	79.3
+ReplaceCap	**37.9**	**94.5**	29.6	80.1
+ReplaceLMCap	37.5	93.4	**31.2**	**83.2**

Table 3.2: Comparing caption replacement with LM generation to ReplaceCap on Flickr30K. Both methods improve over baseline for BLIP and BEiT-3.

Human Study: Errors made by text-to-image generation models To assess the quality of the generated images and their alignment with human judgments, we perform a human study to evaluate the errors present in the synthesized images. This will serve to better understand any shortcomings with the ReplaceImg curation that is not captured by automatic evaluation measures.

We first ranked synthesized images by model loss from the 1K images in the COCO validation set. We then sampled a subset for human annotation using the top and bottom 50 images based on their loss using our fine-tuned captioning model. These images are uniformly divided into 5 sets, each containing 20 images with equal number of the high loss



(a) Distribution of text-to-image generation errors.

(b) Human evaluation versus CLIPScore.

Figure 3.8: Results of the human study of the errors made by the Stable Diffusion model in 100 images. The images used in the study were chosen to represent either low or high model loss. (a) Histogram of the number of errors annotated in each category. The most frequently occurring annotations concern weird deformations in the expected objects or humans. (b) Relationship between average number of identified errors by human annotations for each synthesized image and its captioning loss with regard to original captions. More errors are identified in images of higher loss. However, CLIPScore appears to fail in validating qualities of the synthesized images, as the score ranges are almost identical for samples that contain more errors.

ones and the low loss ones. The data was annotated by 12 people, members of a university research lab with a basic understanding of text-to-image generation but no knowledge of the bi-modal distribution of images. The annotators were asked to categorize the errors in the synthesized images, given both the image and the reference sentences that were used to generate the images. Each participant annotated one set images.

Starting from the categories defined by van Miltenburg and Elliott (2017), we defined 25 error categories including color, number mismatches, and errors related to people and objects in the images. Please see the user interface and more details in the Appendix 3.9.1. We analyze the human judgements for the images that have at least three annotations, yielding 74 unique images.

As shown in Figure 3.8a, the most common problem of the synthesized images are that they often generate weird face or body parts, which makes the images less natural or pleasant. The text-to-image generation model is also weak at generating the correct number of people or objects. From Figure 3.8b we confirm the quality of our collected annotations that high loss figures often contain more errors on average. Furthermore, we

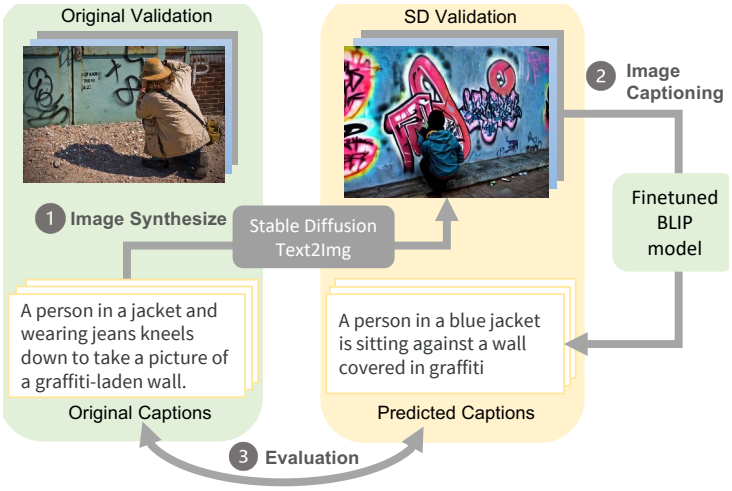


Figure 3.9: Round-trip captioning evaluation.

note that CLIPScore is insensitive to these types of errors, indicating its limited capability of evaluating quality of generated images. Additional examples can be found in Figure 3.13 in the Appendix.

3.7 Further Analysis

With the human study revealing the failure modes of the text-to-image model, we now provide insights on various techniques that are proved useful for improving image relevance in curating the image captioning datasets.

Round-trip captioning evaluation Most previous work in text-to-image generation uses image-oriented measures like FID (Heusel et al., 2017) or CLIPScore (Hessel et al., 2021). However, these measures are not suitable for our purpose as they are claimed to lack alignment with perceptual quality (Saharia et al., 2022). We also found that CLIPScore cannot distinguish between low- and high-loss samples in captioning (Figure 3.8).

Alternatively, similar to Hong et al. (2018), we use a fixed model to generate captions for synthesized images and then compare them to original captions in a three-step process (Figure 3.9): (1) Generating images from validation set captions; (2) Predicting captions for the generated images using a strong image-captioning model; here we use BLIP fine-tuned on the COCO dataset but any other strong captioning model could be used instead. (3) Comparing the predicted captions with the original captions. The assumption is that if the generated images are of similar quality to the originals, the resulting captions will also be similar.

Model	FT	Prompt	B	C	M
Upper-bound			37.6	27.2	57.1
SD 1.5	-	concat	31.0	24.7	52.5
SD 1.5	-	+ styler	30.8	24.2	52.5
SD 1.5	F	+ styler	**33.5**	**25.0**	**53.5**
SD 1.5	F	SBERT + styler	30.6	24.1	52.0
SD 2.0	-	concat + styler	31.2	24.8	52.0

Table 3.3: Round-trip captioning evaluation on Flickr30K with different Stable Diffusion models, prompts, and fine-tuning. F indicates that the model is finetuned. We report **B**LEU, **C**IDEr, **M**eteor.

Ablation on text-to-image variants Evaluating with round-trip captioning, we conduct an ablation study on variants of text-to-image generation models. Table 3.3 summarizes the evaluation results on the Flickr30K dataset. Specifically, we experiment with different versions of the Stable Diffusion models; prompt the diffusion models with various approaches (Section 3.3.2); and compare the generation performance between the finetuned text-to-image model and the pretrained ones. The results show that Stable Diffusion v1.5 finetuned on COCO outperforms the other variants, when prompted with the concatenation of all five captions, with the addition of the styler. For the details of the model variants, please refer to Appendix 3.9.2.

3.8 Conclusion

In this paper, we have shown a simple, yet effective, data curation framework that can improve the performance of image captioning models. We investigated three approaches to data curation that dynamically update the training dataset based on high-loss image-caption samples. The methods involved either removing a sample, replacing the caption in a sample, or generating a new image from existing captions. Experimental results on the Flickr30K and MS COCO datasets show the effectiveness of these approaches to data curation without increasing the total size of the training dataset. A deeper analysis of the images synthesized by the text-to-image model shows frequent errors on generating objects of a certain amount or color, and struggles with human body features. A human evaluation of the errors in those images shows a clear difference in images with high or low losses.

In the future, we expect that better text-to-image generation models will lead to further improvements from using synthesized images to train image captioning models. From our insights in Appendix 3.9.4, there is also significant promise on building a hybrid model

combining different curation methods. We believe that a more sophisticated learning scheme leveraging multiple methods will offer more flexibility when curating the dataset. We plan on verifying whether these findings extend to other image captioning models. Moreover, we are also interested in applying the same framework to other multimodal tasks, especially those with under-complete datasets that cannot comprehensively cover the distributional space due to the cost of crowd-sourcing enough data, e.g. visual question answering, or visually-grounded dialog.

Limitations

As Nguyen et al. (2023) has successfully improved the quality of the pretraining dataset by using an state-of-the-art BLIP-2 model to generate better captions, we would expect that our curation strategies to be scaled and adapted also to vision-language pretraining, which however is limited by research resources and therefore not explored in the scope of this paper. Currently our data curation methods also rely on state-of-the art pretrained models for both image understanding and text-to-image generation.

In our study, we explore how the application of various curation approaches impacts the downstream image captioning performance under different curation ratios. While we predefine the curation ratio for our experiments in this paper, it is desirable for curation methods to be more readily applicable if the curation ratio can be automatically determined.

Moreover, while we take an online approach to data curation, our current approach is upper bounded in speed and performance of the text-to-image generation model. This might be a large bottle neck for adapting the strategy for more complicated vision-and-language tasks.

Ethics Statement

Text-to-image generation is controversial in the broader AI and ethics community(Carlini et al., 2023). For example, it can generate images according to gender or racial stereotypes, which may prove harmful to members of those communities (Li et al., 2022c). While have not yet been observed in the vision-language domain, Shumailov et al. (2023) provide evidence that the use of synthetic data from generative models like large language models can introduce a potential risk of data quality degradation.

In this paper, we use text-to-image to improve the quality of an image captioning model, given a specific set of crowd-sourced captions. Those captions may themselves contain harmful stereotypes that would become more prevalent in our dynamically updated training datasets. As we dynamically update the model with new images based on loss

values, we remove the water-marker in our generated images to prevent information leak to the model. Use of the synthesized images will strictly follow community guidelines.

While developing our curation methods that involve text-to-image generation for image replacement, we employed the stable-diffusion v1.5 model (Rombach et al., 2022), which was trained on the LAION-5B dataset. We note that we were unaware of any investigation into illegal material in the dataset (Thiel, 2023). Hence, we emphasize that our proposed framework is compatible with any other text-to-image models trained on more reliable datasets. Taking this in to consideration, we encourage researchers to explore and apply alternative text-to-image models when incorporating the curation techniques in their future work.

Acknowledgement

We thank Jiaang Li, Lei Li, and the CoAStal and LAMP groups for feedback. Wenyan Li is supported by the Lundbeck Foundation (BrainDrugs grant: R279-2018-1145) and by Innovation Fund Denmark in the context of AI4Xray project. Jonas F. Lotz is funded by the ROCKWOOL Foundation (grant 1242). This work was supported by a research grant (VIL53122) from VILLUM FONDEN.

3.9 Appendix

3.9.1 User interface for human study on categorizing text-to-image generation errors

Our user interface is shown in Figure 3.10. Annotators were asked to tick boxes of errors that they found in the given synthesized images.

The error categories include:

- **People**: age, gender, type of clothing, color of clothing, weird face, weird body
- **Main object**: wrong, similar, inexistent, extra, weird
- **Other objects**: wrong, similar, inexistent, extra, weird
- **General**: stance, activity, position, number, inconsistent references, scene/event/location, text, color, generally unrelated

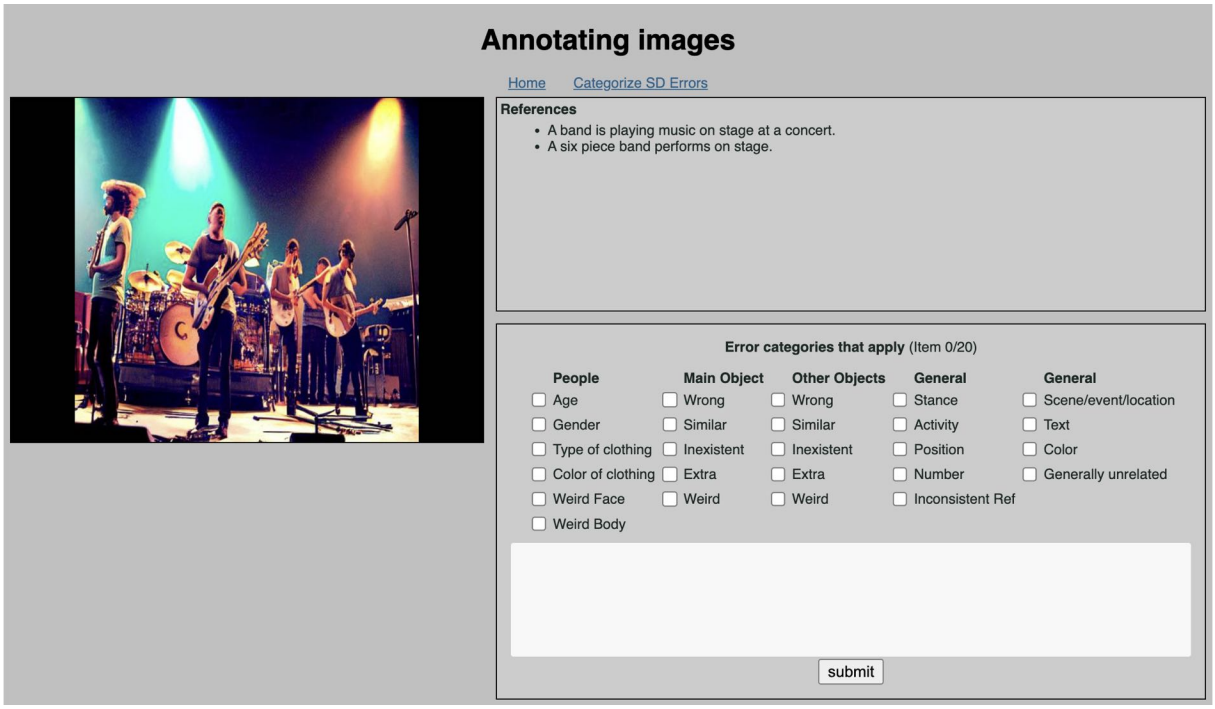


Figure 3.10: Annotation interface for categorizing SD errors.

3.9.2 Prompting approaches for text-to-image generation

Figure 3.11 illustrates the different approaches that we use to prompt the text-to-image generation model. We manually design the styler by inspecting the generated visual examples.

3.9.3 Generating alternative captions with XGLM

We follow the prompt template used in (Ramos et al., 2023b) to obtain LM-generated captions, i.e. "I am an intelligent image captioning bot. Similar images have the following captions: `<captions>` A creative short caption I can generate to describe this image is: `<generation>`". Here we used four ground truth captions as `<captions>` and the other one in `<generation>` for a image to build three-shot examples as the prompt. We used the 'facebook/xglm-2.9B' model which is available on HuggingFace (Wolf et al., 2019). We set the maximum generation length to 30 tokens with number of beams of 5 to prevent from generating repeated tokens.

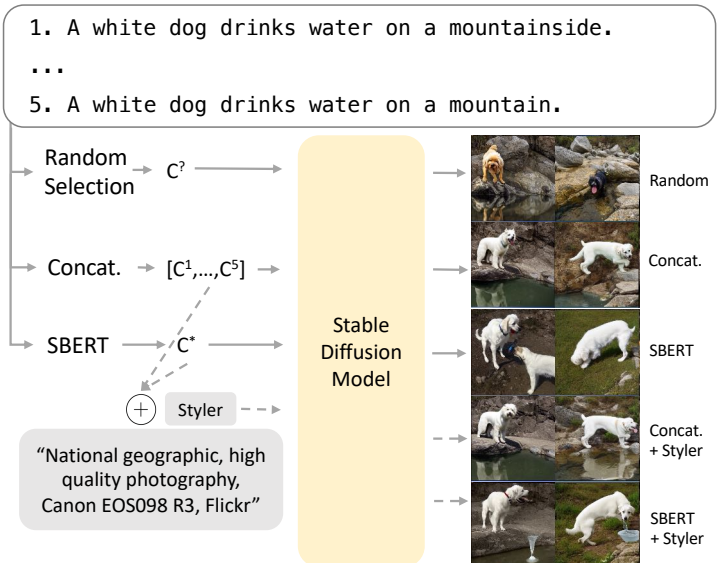


Figure 3.11: Different prompting strategies for synthetic image generation with text-to-image generation and representative examples. Based on our Round-trip Captioning Evaluation, prompting with the concatenated captions and the styler generates the best images for the task.

3.9.4 Combining multiple curation methods

In our pursuit to assess the efficacy of a hybrid model incorporating multiple curation methods, we experiment on the Flickr30K dataset with BEiT-3 as an initial attempt. For the combining strategy, we selected the two most effective methods on the dataset, namely Remove and ReplaceImg. After each training epoch, we curated the training samples by eliminating one half of the top loss samples while substituting the images of the remaining half. Here we curate on the samples with a loss that exceeded two standard deviations from the mean. Our experiment achieves a CIDEr score of 83.8 and a BLEU4 score of 32.8, surpassing previous single curation performance on the dataset. We believe that the hybrid curation approach would yield greater benefits with more sophisticated combining strategies, which we leave for future work.

3.9.5 High-loss training samples

In Figure 3.12, we visualize the high loss training samples in the COCO dataset after the first epoch of finetuning. These samples are target of our curation techniques. Compared to the average caption length of 11 words, the top samples all have very long captions of around 30 words, making it difficult for the model to learn. In the following finetuning epochs, we curate on these samples by either removing the text-image pairs completely (Remove),

Image	Caption	Length	Loss
	a picture of a clearly disrespectful person littered, abused alcohol, didn't flush their bad choices, and worst of all, let old glory touch a bathroom floor	26	213.24
	a picture of a rain-wet street view with lots of bike riders, rimmed with buildings that seem to bunch up and fight for space might look gray and unprepossessing, but doesn't, in part	33	200.14
	a picture of the scene shows outdoors, furthest to closest, shrubbery than a playing field with at least two uniformed and young players, and closest, a blue fence, and a long bench with	33	200.02
	a picture of it is outdoors, the exterior of a low roofed domicile, where a tiny grove of slender tropical trees makes a lean-to for super-modern blue and white motorcycle	30	199.90
	a picture of while a purple/blue sky with what looks like a kite or a loose para-sail floating in it covers most of a distance shot, the bottommost part shows grassy side banks	33	197.36

Figure 3.12: High loss training samples in COCO after the first epoch, ranked by loss in descending order. The top samples all have very long captions around 30 words, compared to the mean of 11 words of the datasets.

replacing the caption (ReplaceCap), or replacing the image with a synthesized unseen image (ReplaceImg).

3.9.6 Examples of synthesized images

In Figure 3.13, we show examples of synthesized images from the text-to-image model that are of high losses and low losses, alongside with the human annotations regarding errors identified from these images.

Image	Caption	CLIPScore	Loss	Categorized Errors
	A picture of two women with one in lacy white dress with handbag and leggings and the other with a tall red hat, black mid-dress, and frame like plastic dress on top.	84.1	181.0	type/color of clothing, color-clothing, weird-face
	A pedicab driver waiting on his bike.	89.3	169.2	weird-main-object, weird-other-object, weird-body-parts, stance
	A man in a black suit with tie and corsage smiles at a girl who smiles back, both are sitting at a table at a semi formal event such as a wedding or reunion.	77.6	163.5	color-clothing, weird-body-parts, wrong-main-object, scene/event/location
	Two men are playing guitars and one man is singing into a microphone on a stage with the spotlight on them.	74.7	26.0	weird-face, weird-body-parts, weird-main-object, weird-other-object
	There a several people in a dark bar-type room, including one girl on a stool.	84.9	26.5	number, weird-face, weird-main-object, weird-body-parts
	Many children are playing and swimming in the water.	78.2	26.9	weird-face, weird-body-parts

Figure 3.13: Examples of synthesized images that are of high losses (top) and examples of synthesized images that are of low losses (bottom). Human annotations show that consistent error types have been recognized for the high loss samples while CLIPScore fails to align with human judgement. The low loss synthesized images are visually less complicated than the higher loss ones, but can still often look weird and contain errors in color or objects.

Chapter 4

FoodieQA: A Multimodal Dataset for Fine-Grained Understanding of Chinese Food Culture

Abstract

Food is a rich and varied dimension of cultural heritage, crucial to both individuals and social groups. To bridge the gap in the literature on the often-overlooked regional diversity in this domain, we introduce FoodieQA, a manually curated, fine-grained image-text dataset capturing the intricate features of food cultures across various regions in China. We evaluate vision–language Models (VLMs) and large language models (LLMs) on newly collected, unseen food images and corresponding questions. FoodieQA comprises three multiple-choice question-answering tasks where models need to answer questions based on multiple images, a single image, and text-only descriptions, respectively. While LLMs excel at text-based question answering, surpassing human accuracy, the open-weights VLMs still fall short by 41% on multi-image and 21% on single-image VQA tasks, although closed-weights models perform closer to human levels (within 10%). Our findings highlight that understanding food and its cultural implications remains a challenging and under-explored direction.

4.1 Introduction

One of the most popular dishes in China is *hotpot*, which comes in many varieties, as shown in Figure 4.1: Beijing is renowned for its mutton hotpot served with a traditional copper pot (铜锅涮羊肉 tóng guō shuàn yáng ròu). Guangdong province is home to a famous porridge-based hotpot (粥底 zhōu dǐ

Figure 4.1: An example of regional food differences in referring to *hotpot* in China. The depicted soups and dishware visually reflect the ingredients, flavors, and traditions of these regions: Beijing in the north, Sichuan in the southwest, and Guangdong in the south coast.

huǒ guō
火锅), while its coastal region of Chaoshan is known for beef hotpot (
cháoshàn niú ròu huǒ guō
潮汕牛肉火锅). The hotpot varieties from Sichuan and Chongqing are celebrated for their flavorful broths, with chili peppers and Sichuan peppercorns that create a unique numbing-spicy sensation. The variation among regional cultures within a country highlights the challenges that language models face in understanding cultural knowledge and context-specific information in the food domain.

Existing datasets and models that focus on food and culinary practices primarily concentrate on tasks such as food recognition, recipe generation, food knowledge probing or recipe-related question answering (Chen et al., 2017; Cao et al., 2024a; Zhou et al., 2024; Yagcioglu et al., 2018). However, they often take a coarse view, conflating country, culture and language. Important regional cultural differences remain under-studied (Palta and Rudinger, 2023).

We introduce **FoodieQA**, a manually curated set of multimodal test questions designed to probe fine-grained cultural awareness with a focus on the food domain. Our dataset targets two under-explored directions: regional cultural diversity within a country and challenging fine-grained vision-language understanding in the culinary domain.

[1]We only evaluate TextQA in Chinese to prevent bias introduced through translating dish names. The English translation is only for illustration purpose.

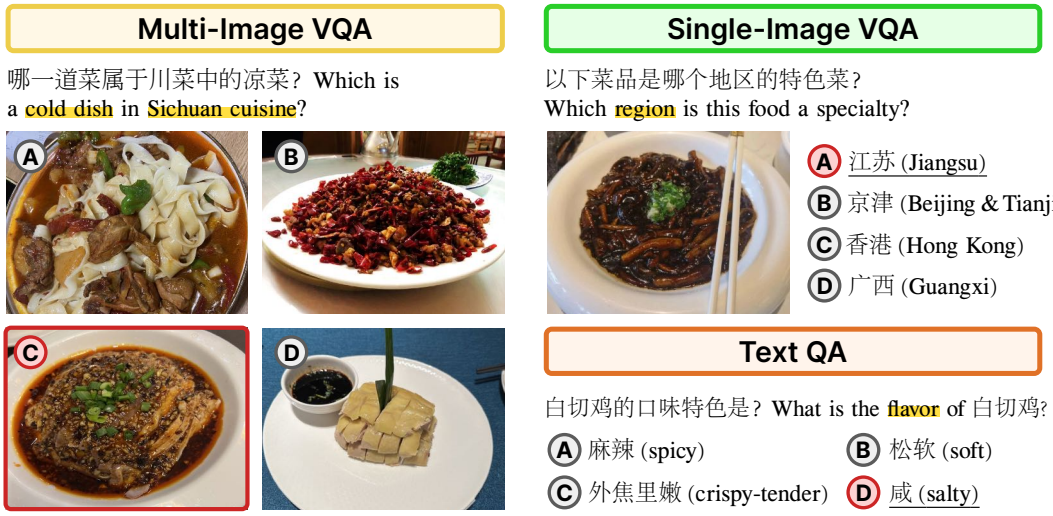


Figure 4.2: The tasks in FoodieQA evaluate food culture understanding from three perspectives. *Multi-image VQA* requires the ability to compare multiple images, similar to how humans browse a restaurant menu. *Single-image VQA* assesses whether models can use visual information to better understand food culture. *Text-based* questions probe model performance without multimodal data.[1]Fine-grained attributes that the questions focus on are highlighted.

To build a regionally diverse dataset, we gather dishes and images selected by native Chinese speakers from various regions, covering 14 distinct cuisine types across China. To ensure the images used for benchmarking are fresh and have no chance of leaking into the pretraining data of VLMs, we collect images uploaded by local people, which are not publicly available online. We then define multiple attributes associated with the dishes and have native Chinese annotators create multiple-choice questions based on their expertise. Our dataset includes both vision-based question answering and text-based question answering tasks, as illustrated in Figure 5.1.

We benchmark a series of state-of-the-art models, including seven LLMs and eight VLMs, on the Foodie dataset using zero-shot evaluation. By comparing their performance to human accuracy, we highlight the gap between open-weights and closed-weights models and demonstrate their limitations in understanding Chinese regional food culture. Additionally, we compare the performance of bilingual models trained on both Chinese and English datasets to English-focused models, revealing biases in their understanding of region-specific food culture and the language of the questions. Finally, our analysis shows that visual information improves the performance of VLMs compared to text-only inputs, although some models struggle with identifying dishes from images

4.2 Related Work

Multilingual Multimodal Datasets Multimodal systems are typically evaluated on English due to the widespread availability of English-language datasets. However, there are some examples of research on training and evaluating models beyond English for image captioning (Elliott et al., 2016), image–sentence retrieval (Srinivasan et al., 2021), visual reasoning (Liu et al., 2021a), and question-answering (Pfeiffer et al., 2022). This paper focuses on Chinese visual question answering, with fine-grained attributes in the food domain.

Food Datasets In recent years, most food datasets have been designed for food image classification (Chen et al., 2017), food captioning (Ma et al., 2023), and recipe-focused generation and question answering (Yagcioglu et al., 2018; Min et al., 2018; Liu et al., 2022). For culture knowledge probing in the food domain, some of the recent datasets span multiple countries and include broad cultural or regional metadata (Min et al., 2018; Ma et al., 2023; Romero et al., 2024). However, they often use country as a proxy for culture, such as the country of origin for the food. For example, Palta and Rudinger (2023) introduced a test set to probe culinary cultural biases by considering US and non-US traditions, Zhou et al. (2024) construct a multicultural, multilingual dataset focusing on culinary knowledge, and Cao et al. (2024a) focuses on recipe transfer between Chinese and English. Investigating cultural differences within a country remains an under-explored area (Palta and Rudinger, 2023).

Fine-grained Vision-Language Understanding Bugliarello et al. (2023) quantified the fine-grained vision-language understanding capabilities in existing models, focusing on aspects within the general domain. Later works focus on the culture understanding in VLMs (Liu et al., 2023b; Cao et al., 2024b). However, current fine-grained VL datasets (Zhang et al., 2021; Parcalabescu et al., 2022; Thrush et al., 2022; Hendricks and Nematzadeh, 2021) are often framed as binary classification tasks, which limits their difficulty. Concurrently with our work, Romero et al. (2024) and Nayak et al. (2024) have created culturally-diverse question-answering datasets across multiple countries. Our multi-choice vision question answering dataset that focuses on Chinese regional differences aims to advance the boundaries of fine-grained understanding in the context of food and culture.

4.3 FoodieQA: Dataset Annotation

China, with its expansive territory and long history, has cultivated rich and diverse food culture and traditions. Focusing on regional food culture differences, our dataset collection contains five distinct phases. 1) selection of cuisine types inside China; 2) collection of

Figure 4.3: Geographical distribution of cuisine types.[2]

private images; 3) individual dish annotation; 4) visual question formulation; 5) text question formulation.

4.3.1 Selection of Cuisine Types

The well-recognized "eight major cuisines" in China are Sichuan (川菜), Guangdong (i.e., Cantonese, 粤菜), Shandong (鲁菜), Jiangsu (苏菜), Zhejiang (浙菜), Fujian (闽菜), Hunan (湘菜), Anhui (徽菜) cuisines (Zhang and Ma, 2020). This categorization is based on historical, cultural, and geographical factors that have influenced the development of distinct cooking styles and flavors in different regions of the country. For a better geographical coverage, we extend the eight cuisine types to additionally include Northwest (西北菜), Northeast (东北菜), Xinjiang (新疆菜), Jiangxi (赣菜) and, Mongolian cuisines (内蒙古菜) in this study. This results in 14 types (Figure 4.3) in total, for which we collect dish images and annotations.

4.3.2 Collection of Images

To ensure that the images are not used in the pretraining of existing models and contaminating evaluation, we designed and distributed a survey for Chinese locals to upload

[2]We omit the Islands of the South China Sea in the figure for visualization simplicity.

Figure 4.4: Meta-info annotation for local specialty.

their own dish images (Figure 4.11).[3] We provide detailed guidelines for image uploading, specifying that: (1) the image should be clear, with a single dish as the focal point in the center; (2) participants should select the cuisine type of the dish from our list or specify it if it is not listed; (3) participants should provide the specific name of the dish, e.g., "mapo tofu (麻婆豆腐)" instead of "tofu (豆腐)"; (4) participants should indicate where the dish was served in their image, choosing from options such as cooked at home, restaurant, canteen, or delivery; (5) participants need to grant us permission to use the image for research purposes and confirm the image is not publicly available online, i.e., it has neither been downloaded from nor uploaded to the web or social media. In other words, the images we collected only existed on their phones or cameras. The uploaded images genuinely represent the locals' daily diet and culinary experiences, showcasing dishes that are currently popular. We manually filter out 102 images that are blurry, have the dish off-center, or show a mismatch between the dish and the image.

4.3.3 Local Specialty Annotation

We also gather text annotations of representative local specialties for each cuisine type on our list. Annotators are asked to collect meta information for representative local dishes for each cuisine type, based on their life experience and knowledge obtained from the web. These meta-fields provide information beyond recipes, offering insights into how the food looks and tastes when people are eating it. An example is provided in Figure 4.4.

The 17 meta-info fields cover the appearance, taste, and culinary attributes of a

[3]The survey is distributed through WeChat and Douban.

dish. They include the food category, dish name, alternative names, main ingredient, characteristics of the main ingredient, three other key ingredients, dish flavor, presentation style, dish color, serving temperature (cold or warm), dishware used, region and province of origin, cuisine type, three primary cooking techniques, eating habits (if any), and reference links.

The annotation is done by eight native Chinese speakers, including five PhD students and three postdoctoral researchers from various provinces in China.[4] During the annotation process, we ensure that all collected data is either annotated or verified by individuals familiar with the local context. Specifically, annotators are assigned as follows: 1) They are asked to annotate local specialties for the cuisine types from their hometowns, guaranteeing that the annotations are provided by locals. 2) If a local annotator can not be found for a specific cuisine type, annotators are requested to seek assistance from friends who are from the respective region to verify or correct the metadata obtained from the web. Annotations in the following sections are conducted by the same annotators, if not mentioned otherwise.

4.3.4 Visual Question Answering Annotation

One major consideration for vision-language understanding is that models can rely on language priors, consequently neglecting visual information (Goyal et al., 2017; Zhang et al., 2016). This underscores the importance of formulating visual questions in such a way that they can only be answered by examining visual features, rather than relying on text priors. Based on the number of images used as inputs, we formulate both multi-image VQA questions and single-image VQA questions.

4.3.4.1 Multi-image VQA

Multi-image VQA requires the ability to compare detailed visual features from multiple images, similar to how humans browse a restaurant menu.

Question formulation We ask the annotators to write challenging questions that require: (1) looking at the dish images to answer, (2) thinking beyond merely recognizing the dish and questions that may require multi-hop reasoning, (3) asking diverse questions that belong to a diverse set of question types such as food type, flavor, color, expense, amount, and etc., (4) only one image is the correct answer to the question. The multi-image VQA questions are written by five native speakers from five different regions in China.

We organize the collected images into 28 groups based on cuisine types and food categories, as outlined in Section 4.3.2. This allows annotators to write questions sequentially for related images extracted from the same group. Each annotator is asked to write

[4]The annotators are from Sichuan, Shaanxi, Guangdong, Jiangsu, Jiangxi, Shandong, and Chongqing.

two–three questions, given a four-image group. We note that in order to avoid the bias from language priors, dish names corresponding to the images are not presented. The user interface that we use for annotation is shown in Figure 4.12.

Question verification Once the questions and answers for the multi-image multiple-choice questions are collected, we verify the questions by asking the annotators (who did not create the questions) to answer them. If a question does not meet our defined criteria, annotators are instructed to flag it as a "bad question". Through this process, 87 questions were discarded. Additionally, when answering the questions, annotators are required to provide the rationale they use to reach the answer, as well as judge whether the question requires multi-hop reasoning. The user interface that we use for verification is shown in Figure 4.13. Each question is verified by two annotators, and we exclude the questions that do not have full agreement.

4.3.4.2 Single-Image VQA

Besides using images as multiple-choice answer options, we also ask diverse fine-grained questions about various aspects of a dish based on its meta-information (collected in Section 4.3.3). We identify dishes that have both meta-information annotations and collected images, and then create questions based on the meta-information. As shown in the example in Figure 5.1, the dish name is intentionally omitted from the questions to ensure they can only be answered by examining the visual features.

Question formulation We adopt a template-based approach, where a question about the same meta-field is asked multiple times, varying factors like the image of the dish, while the answer options are carefully selected from the wrong candidates in the meta-field to ensure that only one answer is correct. The single-image VQA questions are generated using a rule-based method, followed by thorough human verification that is similar to the multi-image VQA verification process. Please see details in the Appendix 4.7.1.

Question verification Similar to verification for the multi-image VQA questions, annotators are asked to answer the question given the text query and the corresponding image, and raise a "bad question" flag to filter out questions that does not satisfy the criteria. 88 questions were discarded as bad. Note that the name of the dish is not revealed in the text question so that the question needs to be answered based on visual information. Annotators are asked to write "I don't know" in the rationale and randomly guess an answer if they think the question is beyond their knowledge.

4.3.5 Text Question Answering Annotation

We formulate the text-based questions by combining human annotations and rule-based generation. Similar to the single-image VQA approach described in Section 4.3.4.2, we generated questions and multiple-choice answer options based on the meta-information fields. However, instead of using the dish image, we included the dish name directly in the question. The questions are formulated using templates, where only the dish names and meta-fields are varied. A same human verification process to single-image question answering is included. 135 bad questions were discarded. Notice that annotators were asked to answer the questions based on their knowledge without using search engines, this makes the task challenging as it would be hard for one to answer questions about unfamiliar foods and regions without any other available information besides names of the food.

4.3.6 Human Validation

In Table 4.1, we calculate human accuracy and inter-annotator agreement scores based on human-verified questions, excluding those identified as bad questions. For the single-image VQA and text QA questions, given the diverse cultural backgrounds of the human annotators, some questions can be challenging if the required food culture knowledge falls outside an annotator's cultural experience. For those questions, annotators are instructed to indicate "I don't know" and randomly guess an answer, as one might not be familiar with all of the specific dishes or the fourteen cuisine types. These questions are marked as out-of-domain. Considering the randomly selected answers for these out-of-domain questions allow us to obtain lower bound agreement and human accuracy scores.[5] We also report Cohen's Kappa (κ) and human accuracy separately for in-domain questions.

The human validation process involves three postdoctoral researchers and five PhD students who are native Chinese speakers as introduced in Section 4.3.3. Each question is verified and answered by two annotators who were not involved in the question formulation. We retain the out-of-domain questions for calculating human accuracy and later in evaluating model performance, as the lower agreement scores are only due to differences in the annotators' cultural knowledge (Plank, 2022).

4.3.7 Image and Question Distribution

Image statistics We collected 502 images but discarded 113 due to quality control issues. The final dataset of 389 images are distributed across regions in China as shown in

[5]Note that this is the only impact of the randomization. The ground truth label is annotated at an earlier stage of question formulation where the questions and choices are generated using the rule-based method.

Task	Questions	κ	Accuracy
Multi-image VQA	403	.834	.916
Single-image VQA	256	.556	.744
- In-domain	168	.674	.818
Text QA	705	.470	.562
- In-domain	307	.808	.857

Table 4.1: Statistics per task in FoodieQA.

Statitics	Multi-image	Single-image	TextQA
Avg. length	12.9	17.0	14.9
Multi-hop (%)	25.3	73.4	1.6
Question types	14	6	7
Unique Images	389	103	-

Table 4.2: Question statistics.

Figure 4.5. All 389 images are used for multi-image VQA; a subset of 103 images are used for single-image VQA.

Question statistics After human verification, we obtain 403 multi-image VQA questions, where each question needs to be answered with a set of four provided images. Single-image VQA tasks consists of 256 question in total, and text QA consists of 705 questions in total (Table 4.1). A considerable number of the VQA questions require multi-hop reasoning to predict the correct answer. We report the key statistics of the questions in Table 4.2. Please see more details in Appendix 4.7.2.

4.4 Baselines: How Much of a Foodie are the LLMs/VLMs?

We evaluate open-weight and API-based state-of-the-art LLMs and VLMs to probe their culture knowledge in the food domain. We evaluate the models in both Chinese and English for the VQA tasks. The questions are translated to English using the DeepL free API[6] and validated by two PhD students who are Chinese native speakers and fluent in

[6] https://www.deepl.com/en/translator

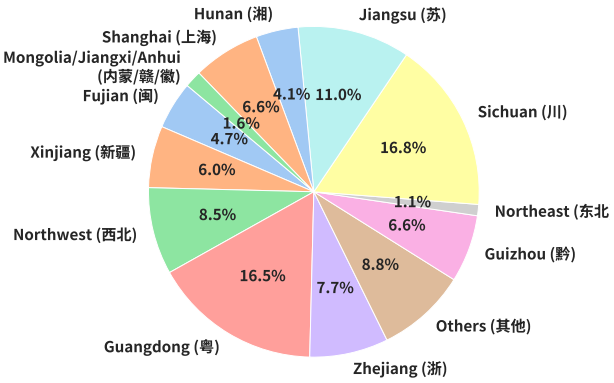


Figure 4.5: Region distribution of collected food images.

English. To avoid bias in translating dish names, we conduct the TextQA task solely in Chinese.

4.4.1 Multi-Image VQA is Difficult

We evaluate the multi-image VQA task using open-weight models that are capable of handling multiple image inputs, including Phi-3-vision-128k-instruct (Abdin et al., 2024), Idefics2-8B (Laurençon et al., 2024), Mantis-8B-Idefics2 (Jiang et al., 2024), and English-Chinese bilingual Qwen-VL-12B (Bai et al., 2023), and Yi-VL 6B and 34B models (AI et al., 2024), as well as API-based models GPT-4V and GPT-4o (Achiam et al., 2023).

We experimented with four different prompts that utilized lists of images and texts or interleaved image-text inputs. Details can be found in Appendix 4.7.4. As shown in Figure 4.6, when compared to the human accuracy of 91.69% in Chinese, the best-performing open-weight model, Idefics2-8B, achieves an accuracy of 50.87%, which is still significantly lower than human performance. This indicates that current state-of-the-art models are still weak at distinguishing differences among food from visual input. This underscores that multi-image understanding, especially in contexts requiring cultural knowledge in the food domain, remains a challenging problem. When evaluating on the translated English questions, model performance decreases for all models except Phi-3-vision.

4.4.2 Single-Image VQA Results

Besides the four open sourced models that we used for multi-image VQA, we also evaluate the bilingually trained (Chinese and English) Yi models (AI et al., 2024) for the single-image VQA task.

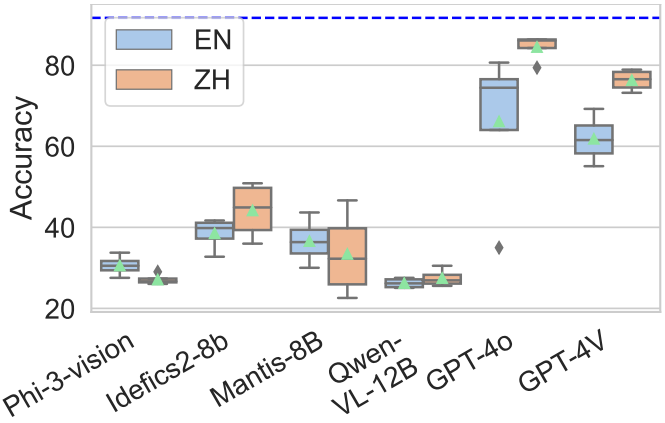


Figure 4.6: Accuracy of multi-image VQA tasks across four different prompts compared to a 91.96% human accuracy in Chinese. Although Idefics2 and Mantis have higher accuracy than other models, they show greater variation across different prompts.

The evaluation accuracy is reported in Table 4.3. Almost every open-weight model performs better on Single-image VQA than Multi-image VQA. We can observe that, for the bilingually trained models, i.e., Qwen-VL and Yi-VL, their performance is better when evaluated in Chinese. However, for the multilingual models, i.e. Phi-3, Idefics2, and Mantis-8B, their performance is better when evaluated in English. The best performing models are the API-based models from OpenAI.

4.4.3 Models are Strong at Text QA

We evaluate text question answering with a series of open-weight models, including Phi-3-medium-4k-instruct (Abdin et al., 2024), Llama3-8B-Chinese (Wang and Zheng, 2024), Mistral-7B-Instruct-v0.3 (Wang and Zheng, 2024), Yi-6B and 34B models (AI et al., 2024), and Qwen2-7B-instruct (qwe, 2024), as well as API-based model GPT-4 (Achiam et al., 2023).

Given that translating dish names is challenging and would likely introduce additional information and unfair comparison, we only evaluate the text questions in Chinese. For example, a famous Sichuan dish "夫妻肺片 (fū qī fèi piàn)" can be translated to "couple's lung slices" if translate word by word, however it would be translated as "Sliced Beef and Ox Tongue in Chilli Sauce" by meaning. While the literal translation makes no sense, translation by meaning would hint the flavor and ingredients that are not included in its original Chinese name.

From Figure 4.7, we see that the Qwen2-7B-instruct model surpasses human performance on the text QA task, where the questions are formulated based on the local specialty annotations in Section 4.3. Since the local specialty annotations are collected and

Evaluation	Multi-image VQA		Single-image VQA	
	ZH	EN	ZH	EN
Human	91.69	77.22†	74.41	46.53†
Phi-3-vision-4.2B	29.03	33.75	42.58	44.53
Idefics2-8B	**50.87**	41.69	46.87	**52.73**
Mantis-8B	46.65	**43.67**	41.80	47.66
Qwen-VL-12B	32.26	27.54	48.83	42.97
Yi-VL-6B	-	-	**49.61**	41.41
Yi-VL-34B	-	-	52.73	48.05
GPT-4V	78.92	69.23	63.67	60.16
GPT-4o	**86.35**	**80.64**	**72.66**	**67.97**

Table 4.3: Comparison of Multi-image and Single-image VQA Performance in Chinese and English. We report the best accuracy from four prompts. † results denote an estimate, calculated over 100 random samples, of human performance on the English Multi-Image and Single-Image VQA from one native speaker with no specialized knowledge of Chinese food culture.

summarized by local representatives, potentially incorporating information from public web resources such as Baidu-Baike, the high performance may be attributed to the inclusion of domain-specific training data.

4.5 Analysis

In this section, we explore which factors are important for fine-grained understanding of Chinese food culture.

Non-public images are crucial for fair evaluation. We incorporate user-uploaded non-public images into our dataset to prevent data contamination during evaluation. To verify the importance of preserving these non-public images for fair evaluation, we compare model performance using web-sourced images instead. Specifically, we manually searched with dish names to obtain web images for 171 out of 256 questions in the Single-image VQA task. As shown in Table 4.4, replacing non-public images with web-sourced dish images made the task easier for baseline models, indicating potential data contamination from web sources. Therefore, the use of non-public images is crucial for ensuring fair evaluation.

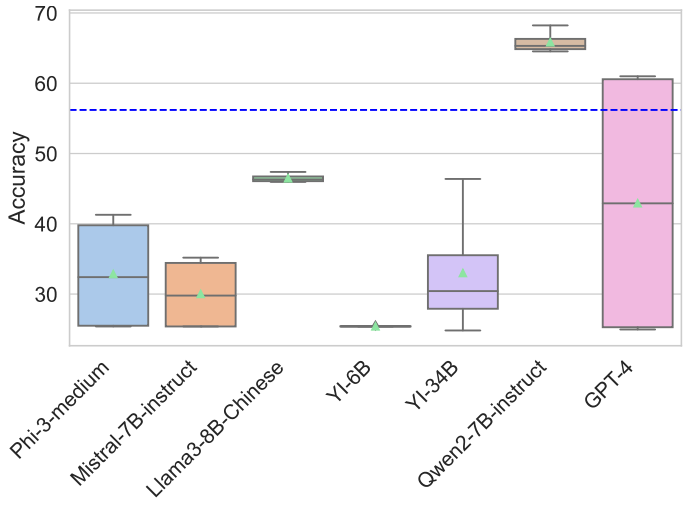


Figure 4.7: Accuracy of text QA across four different prompts. The blue dashed line indicates human accuracy (56.2%).

Model	Non-public images	Web images
Qwen-VL-12B	43.75	**47.95**
Idefics2-8B	45.60	**47.07**
Yi-VL-6B	47.56	**50.88**

Table 4.4: Models obtain higher accuracy when evaluating with web images, which indicates possible data contamination. The accuracy scores are averaged over four prompts.

Visual information helps. In Single-image VQA, the default setting is to query with only dish image without specifying the dish name. We now examine whether the visual information is beneficial using the Idefics2-8B model.[7] Results are shown in Table 4.5, where we investigate two variants: querying the model with only the text question but revealing the dish name, versus providing both the dish image and the dish name. We observe that the Idefics2 model consistently performs better when dish images are available as visual clues. Please see comparison examples in Appendix 4.7.5.

Dish names could be helpful clues for some of the models. As discussed in Section 4.3.7, over 73.4% of single-image questions require multi-hop reasoning, which typically involves identifying the dish and then leveraging related knowledge to answer the questions. To determine whether the identification of the food image and the utilization

[7]We selected this model because it supports text-only inputs, unlike some other models such as the Yi-VL series.

Input	prompt1	prompt2	prompt3	prompt4
Dish name only	28.52	27.73	36.72	37.11
+ dish image	40.23	41.41	40.62	42.19

Table 4.5: Accuracy on two variants of Single-image VQA task, showing that visual information of food images is crucial for Idefics2 to correctly answer the questions.

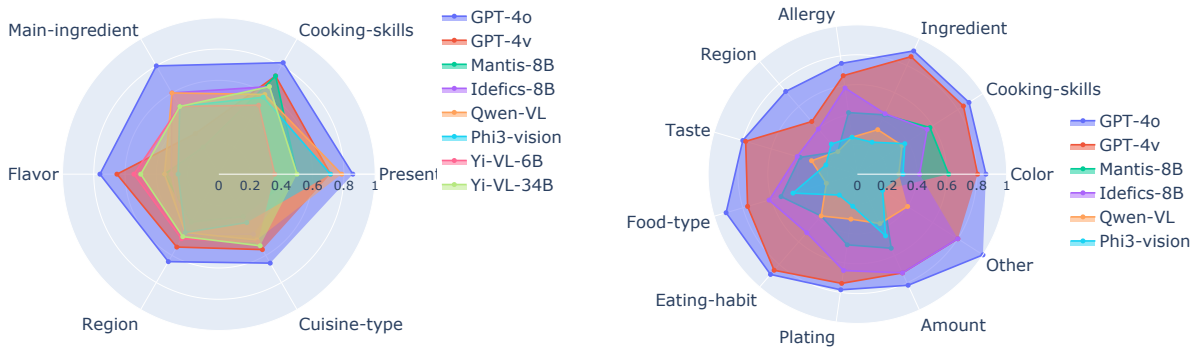
Model	Condition	p1	p2	p3	p4
Yi-VL-6B	Image-only	**49.61**	48.05	47.66	46.09
	+ dish name	73.83	74.61	**76.17**	62.50
Yi-VL-34B	Image-only	50.39	**52.73**	50.78	48.83
	+ dish name	75.39	78.13	**79.30**	75.39
Idefics2-8B	Image-only	44.53	43.75	46.09	**46.87**
	+ dish name	40.23	41.41	40.62	**42.19**

Table 4.6: Accuracy in the Single-image VQA task when dish name is revealed in the questions along with the image or not. While the Yi models benefit greatly from the additional information of the dish name, Idefics2 does not. "p1–4" indicates four different prompt templates.

of visual information are bottlenecks for the models, we compare their performance on single-image VQA when provided with the dish name in the question.

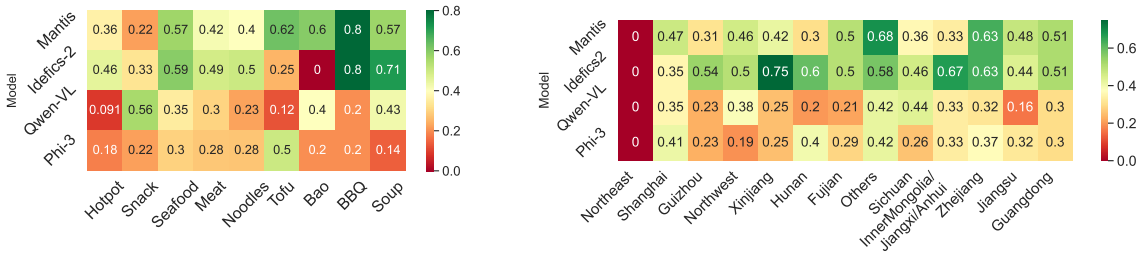
The results in Table 4.6 indicate that while the Yi models significantly benefit from being given both the images and names of the dishes, the Idefics2-8B model does not show the same improvement from this additional information. This indicates that recognizing the dishes could be a possible bottleneck for the Yi series models.

Models are foodies who know cooking better than taste. Figure 4.8a shows the model performance under fine-grained questions attributes on Single- and Multi-image VQA. We observe that all models generally excel at answering questions related to cooking skills and ingredients. The Yi models, in particular, demonstrate a stronger ability to identify the flavors of dishes. Conversely, the Qwen-VL and Phi3-vision models perform well in observing the presentation of food when served but struggle with flavor-related questions. When answering questions based on multiple images, it also holds true that models are generally good at questions regarding cooking skills and the amount of food (Figure 4.8b). However, these models are weak at answering questions related to the region and taste of the dish. Idefics-8B stands out, excelling in most of the fine-grained features



(a) Single-image VQA. (b) Multi-image VQA.

Figure 4.8: Model accuracy on fine-grained question attributes.



(a) Accuracy by food categories. (b) Accuracy by cuisine types.

Figure 4.9: Model accuracy on questions categorized by food categories and cuisine types.

we evaluated.

Favorite food of the models. In Figure 4.9, we compare model performance on multi-image VQA tasks for questions grouped by food categories and cuisine types. This analysis provides insight into how well the models can compare features from images within the same group. The overall best performing model on multi-image VQA tasks excels at questions about BBQ and Xinjiang cuisines, but weak at questions about Shanghai dishes. Another interesting finding is that, despite Sichuan food being one of the most popular cuisines in China, and presumably having more available images and resources online, none of the models excel at answering questions related to this cuisine type.

4.6 Conclusion

We introduce FoodieQA, a multimodal dataset designed to evaluate fine-grained understanding of Chinese food culture through multi-image, single-image, and text-only multiple-choice questions.

Our experiments, which focus on regional cultural differences and detailed visual features, reveal that understanding food and its cultural context remains a complex and under-explored task. We find that comparing food across multiple images—similar to the common scenario of people browsing menus—is particularly challenging. All open-source models underperform human accuracy by more than 40% in this task. This suggests that our dataset offers a more accurate assessment of the suitability of state-of-the-art models for real-world applications in the food domain.

Our analysis of language and prompt templates indicates that models can be sensitive to the language in which questions are asked—bilingually trained Chinese–English models perform better in Chinese, while other multilingual models are stronger in English. We also demonstrate the effectiveness of incorporating visual features compared to text-only settings in this context.

Improved models or methods for understanding food culture may be essential for future progress in the FoodieQA challenge. Looking ahead, we aim to expand the dataset to include dishes from other countries and regions. Following Jacovi et al. (2023), we make our dataset a public benchmark on Huggingface at lyan62/FoodieQA with the CC BY-NC-ND 4.0 License. All of our data annotation and verification tools are freely available for re-use at github.com/lyan62/FoodieQA. We encourage the community to create Foodie datasets for their own language and culture groups.

Limitations

The size of the FoodieQA dataset is limited by the challenge of collecting unseen images from individuals, as it requires them to voluntarily upload images from their phones or cameras. Although we have distributed the survey on two popular Chinese social media platforms, we anticipate that increased social media exposure or collaboration with food industry professionals could facilitate the collection of more images, and contribute to a training dataset for advancing this direction.

Translating Chinese dish names into other languages poses another challenge, as some dish names do not directly relate to their ingredients or cooking methods. Introducing translated dish names could potentially introduce additional information, leading to unfair comparisons among the models. Consequently, we have chosen to experiment solely with Chinese questions for the text-based queries.

We have benchmarked fifteen popular models using our dataset. However, due to

the rapid advancements in the field, it is impossible to benchmark all trending models continuously. We hope our dataset will inspire future researchers to develop similar Foodie datasets for their own regions and languages, thereby guiding LLMs and VLMs towards a better understanding of regional food cultures.

Acknowledgements

We are grateful to the volunteers for their generous contributions and efforts in providing high-quality food images that support our research. We extend our gratitude to Xi Liu, Yihe Zhang, Yu Sun, Yueyin Xu, Gefan Yang, Shixiong Wang, Penglong Ma, Daiwei Wang, Bo Cui, Yu Dong, Jinming Hu, Yufei Lin, Zhongsheng Huang, Xinyu Shi, Yan Shi, and Yue Shi for serving as local experts. Their efforts in verifying and correcting the local specialty annotations and providing valuable feedback have been essential in ensuring the annotation's accuracy and completeness. We also thank Fengyuan Liu, Ruixiang Cui, Zhi Zhang, Yu Sun, and many of our friends and family who helped spread the image collection survey on social media for wide regional and group coverage. Special thanks to Jordan Boyd-Graber and Jimmy Lin for providing helpful research advice. Wenyan Li is supported by the Lundbeck Foundation (BrainDrugs grant: R279-2018-1145) and a research grant (VIL53122) from VILLUM FONDEN. Jiaang Li is supported by Carlsberg Research Foundation (grant: CF221432) and the Pioneer Centre for AI, DNRF grant number P1. Li Zhou is supported by Shenzhen Science and Technology Research Fund (JCYJ20220818103001002) and Shenzhen Science and Technology Program (ZDSYS20230626091302006).

4.7 Appendix

4.7.1 Rule-based question formulation

For text-based question answering we develop a rule-based question formulation method. For each question type, we have the meta information from the local specialty annotation (Section 4.3.3). Then we design three to four templates for each of the question type. For example, for questions that ask about cuisine type, our templates include

- <dish>是哪个地区的特色菜? (What region is <dish> a specialty dish of?)
- <dish>是哪个地区的特色美食? (In which region that <dish> is a local specialty?)
- 去哪个地方游玩时应该品尝当地的特色美食<dish>? Which place should you visit to taste the local specialty food <dish>?

Then, we randomly select cuisine types that are not the correct answer to serve as the alternative options. By utilizing different meta fields, we can generate multiple questions for each dish.

For single-image VQA, we associate the questions related to the dish with the corresponding dish image in our collection. We exclude questions of the warm-cold type—those that inquire whether a dish is served hot or cold—since these questions involve different dishes as options and are not suitable for the single-image scenario.

4.7.2 Question type and answer distribution

In Table 4.7, 4.8, and 4.9, we show concrete statistics about distribution of question types in each task. Figure 4.10 illustrates the answer distribution for questions categorized by type. Each horizontal bar independently displays the distribution of the answers regarding to the specific question type.

Question type	**Count**
Cuisine Type	147
Cooking Skills	127
Main Ingredient	70
Region	148
Flavor	117
Present	25
Warm-Cold	71

Table 4.7: Distribution of text QA question types.

Question type	**Count**
Cuisine Type	70
Flavor	46
Region	65
Present	14
Cooking Skills	51
Main Ingredient	10

Table 4.8: Distribution of single-image VQA question types .

4.7.3 Annotation Cost and Compensation

In this work, the annotators are our colleagues who share co-authorship of the paper. This applies to the human annotation and validation process in Section 4.3.3, Section 4.3.4, and

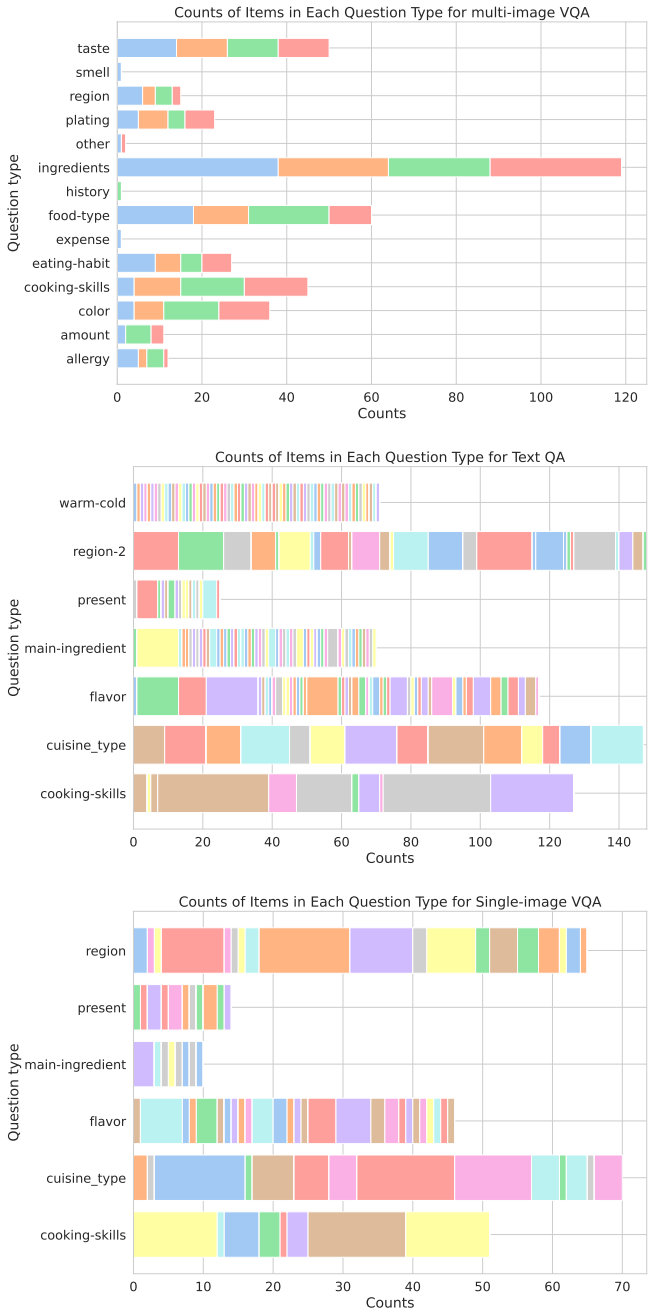


Figure 4.10: Answer distribution for each of the tasks. The questions are categorized by question type. Each color corresponds to a distinct answer, and each horizontal bar displays the distribution of these answers.

Question type	Count
Ingredients	119
Food Type	60
Color	36
Taste	50
Cooking Skills	45
Plating	23
Eating Habit	27
Allergy	12
Region	15
Expense	1
Other	2
Amount	11
Smell	1
History	1

Table 4.9: Distribution of multi-image VQA question types .

Task	Avg time/annotation	Avg time/person
Local specialty collection	11.4 min/dish	10.3 hrs/person
Multi-image VQA question formulation	3.5 min/question	8.0 hrs/person
Multi-image VQA question verification	2.5 min/question	6.7 hrs/person
Single-image VQA verification	3.3 min/question	6.3 hrs/person
TextQA verification	1.2 min/question	5.7 hrs/person

Table 4.10: Average time per annotation and per person for annotation tasks.

Section 4.3.6. The collection of images from private individuals, described in Section 4.3.2, was entirely voluntary and by community effort through the social platforms, WeChat and Douban.

The image collection period takes around one and a half months through the survey. Table 4.10 displays an estimation of the annotation time reported by annotators.

4.7.4 Prompts used for evaluation

Following Durmus et al. (2023) and Wang et al. (2024), we design four prompts for each of the tasks and extract the option letter from the model response. For multi-image VQA, we specifically include prompts that feature both interleaved image and text inputs as well as separate lists of images and texts. Please see examples of the prompts in Table 4.11 and Table 4.12.

Prompt	Content
Prompt 0	<img1><img2><img3><img4> Answer the following question according to the provided four images, they correspond to Option (A), Option (B), Option (C), Option (D). Choose one best answer from the given options. Question: , your answer is: Option (
Prompt 1	Answer the following question according to the provided four images which correspond to Option (A), Option (B), Option (C), Option (D). Choose one best answer from the given options. The options are: <img1>Option (A) <img2>Option (B) <img3>Option (C) <img4>Option (D) Question: <question>, your answer is: Option (
Prompt 2	Answer the following question according to the provided four images, and choose one best answer from the given options. The options are: <img1>Option (A) <img2>Option (B) <img3>Option (C) <img4>Option (D) Question: <question>, your answer is: Option (
Prompt 3	Human: Question <question> The options are: Option (A)<img1> Option (B)<img2> Option (C)<img3> Option (D)<img4> Assistant: If I have to choose one best answer from the given options, the answer is: Option (

Table 4.11: English prompts for zero-shot evaluation for multi-image VQA.

Interface of image collection, annotation and verification tool

In Figure 4.11, we display the survey that we used to collect images. Figure 4.12 and Figure 4.13 show the user interface that annotators use to create questions and verify the questions.

Prompt 1	<img1>,<img2>,<img3>,<img4> 根据以上四张图回答问题，他们分别为图A, 图B, 图C, 图D, 请从给定选项ABCD中选择一个最合适的答案。问题：<question>, 答案为：图
Prompt 2	<img1>,<img2>,<img3>,<img4> 根据以上四张图回答问题, 请从给定选项ABCD中选择一个最合适的答案。问题：<question>, 答案为：图
Prompt 3	根据以下四张图回答问题, 请从给定选项ABCD中选择一个最合适的答案。 <img1>图A <img2>图B <img3>图C <img4>图D 问题：<question>, 答案为：图
Prompt 4	Human: 问题<question>，选项有: 图A<img1> 图B<img2> 图C<img3> 图D<img4> Assistant: 如果从给定选项ABCD中选择一个最合适的答案，答案为：图

Table 4.12: Chinese prompts for zero-shot evaluation for multi-image VQA.

4.7.5 More examples

Examples of the questions in the dataset

See Figure 4.14 for more examples of the questions in the dataset.

Examples of comparing whether the visual information is available

In Figure 4.15, we present examples where visual information, specifically the dish images, proves crucial for the Idefics-2-8B model to accurately answer the questions.

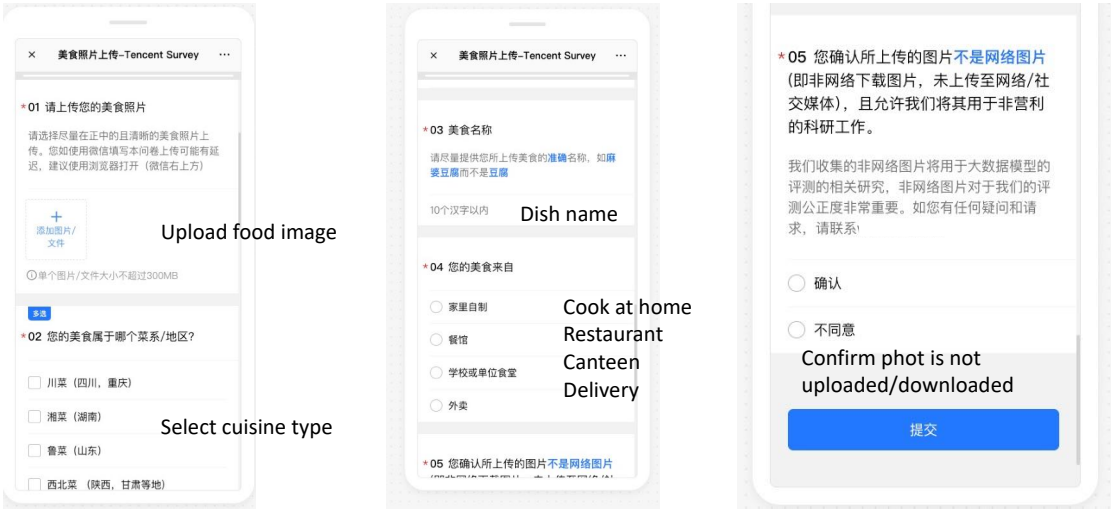


Figure 4.11: Survey interface of image collection

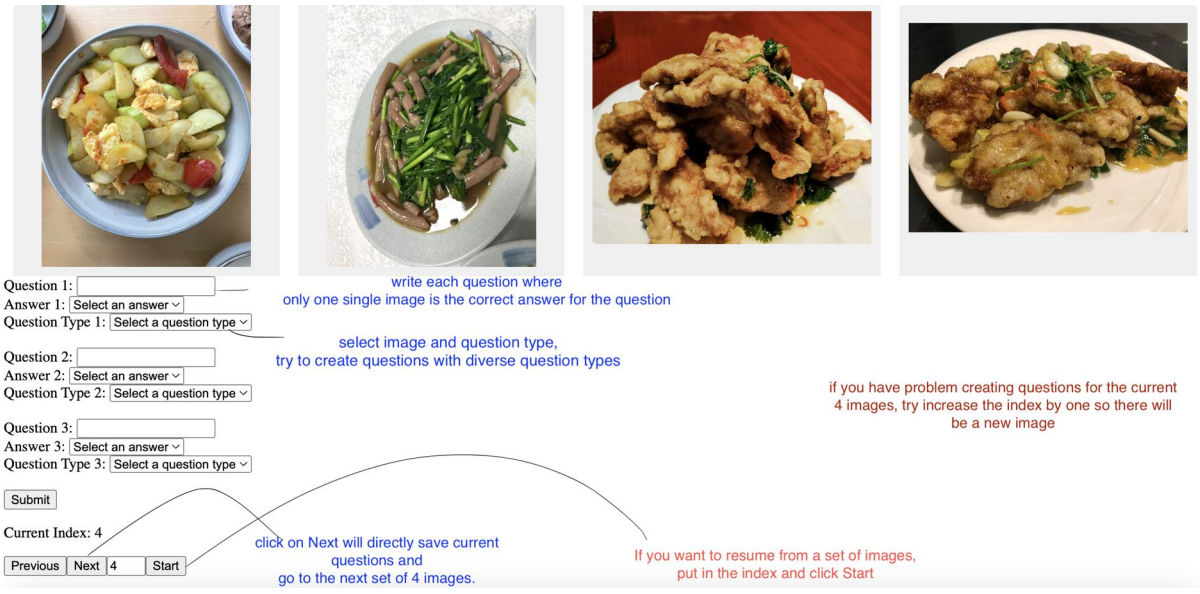


Figure 4.12: Annotation interface of writing questions when presented multiple images.

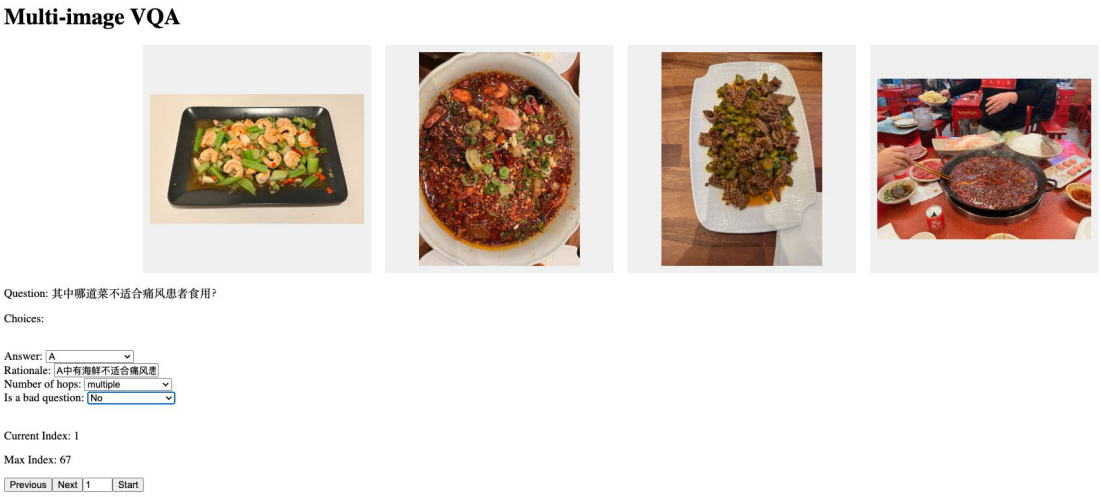


Figure 4.13: Annotation interface of verifying the multi-image multiple-choice questions.

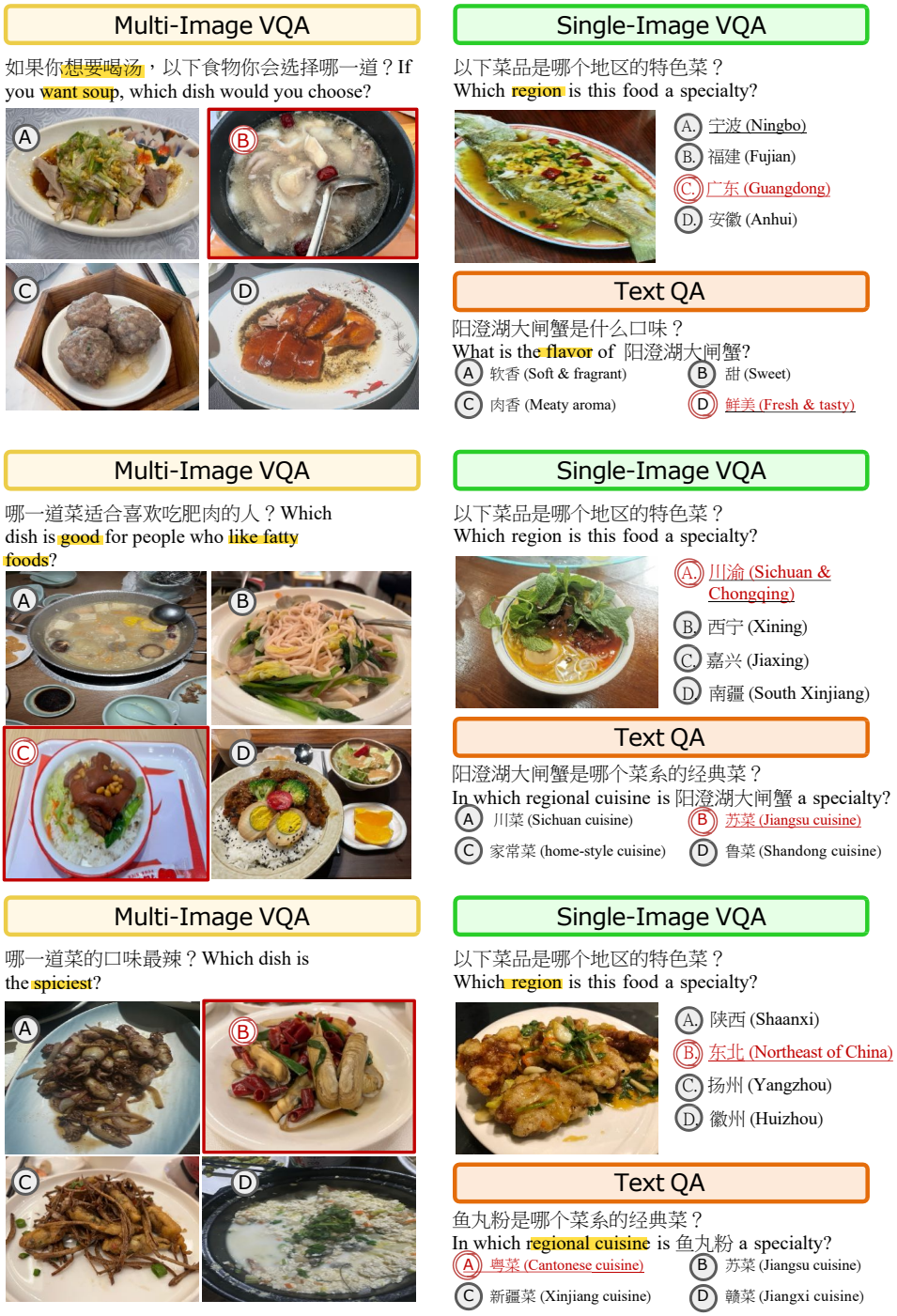


Figure 4.14: More examples in FoodieQA evaluate food culture understanding from three perspectives.

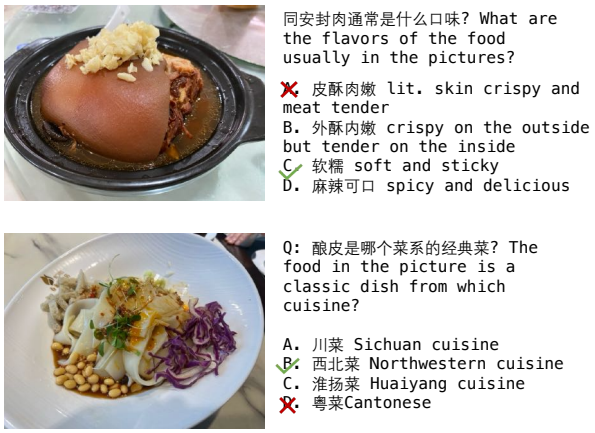


Figure 4.15: Examples where the Idefics-2-8B model correctly answers the question when the image is available but failed when it is not.

Chapter 5

Understanding Retrieval Robustness for Retrieval-Augmented Image Captioning

Abstract

Recent advances in retrieval-augmented models for image captioning highlight the benefit of retrieving related captions for efficient, lightweight models with strong domain-transfer capabilities. While these models demonstrate the success of retrieval augmentation, retrieval models are still far from perfect in practice: the retrieved information can sometimes mislead the model, resulting in incorrect generation and worse performance. In this paper, we analyze the robustness of a retrieval-augmented captioning model SmallCap. Our analysis shows that the model is sensitive to tokens that appear in the majority of the retrieved captions, and the input attribution shows that those tokens are likely copied into the generated output. Given these findings, we propose to train the model by sampling retrieved captions from more diverse sets. This decreases the chance that the model learns to copy majority tokens, and improves both in-domain and cross-domain performance.

5.1 Introduction

Recent retrieval-augmented image captioning models have shown success in strong image captioning performance while reducing model parameters by retrieving related captions for a given image (Ramos et al., 2023c; Sarto et al., 2022; Yang et al., 2023). These models use retrieved information as additional context besides the input image. However, similar to retrieval-augmented language models (Yoran et al., 2023), image captioning models

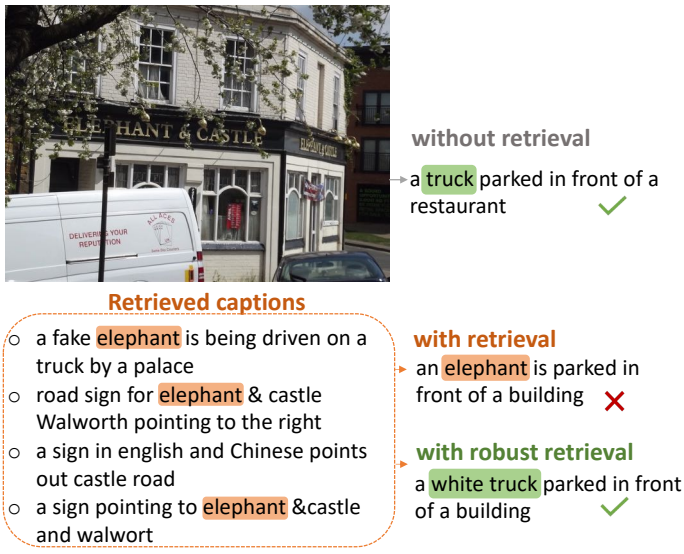


Figure 5.1: Comparison of generated image captions that are predicted without retrieval, misled by retrieval, and predicted with a more retrieval-robust model. The retrieval-augmented model generates the token "elephant", which appears in 3/4 of the retrieved captions.

enhanced with retrieval can sometimes be misled by irrelevant information. For example, in Figure 5.1 the captioning model is misled by the token "elephant" in the retrieved captions, and generates captions that do not match the given image.

For retrieval-augmented language models, Yoran et al. (2023) have studied the cases where retrieval misled the model prediction, and address this problem with a retrieval-robust LLM by continuous training with synthetic data for question answering tasks. However, in their approach, the retrieval system returns only one passage at each step. Considering that LLMs can be sensitive to the order of prompts (Lu et al., 2022), the robustness of using multiple retrieved results has not been fully studied. Evaluating and improving the robustness of retrieval-augmented image captioning models remains under-explored, specifically when the model is augmented with multiple retrieved results.

To bridge this gap in the literature, we study the robustness of the SmallCap retrieval-augmented captioning model (Ramos et al., 2023c). By the definition of retrieval robustness proposed in Yoran et al. (2023), retrieved context should boost model performance when relevant, and should not adversely affect it when irrelevant. We thoroughly examine the robustness of the model with regards to the order of the retrieved captions, and the relevance of the retrieved content. We also present a novel analysis of model behaviour based on *majority voting*, supported by input attribution and attention analyses to investigate how the retrieved tokens influence the model generation. And finally, inspired by Hoang et al.

(2022), we propose to sample the retrieved captions from a larger list during training to prevent the model from overfitting to the top relevant captions. Our evaluation shows improved model robustness and better out-of-domain generalization.

The main findings of this paper are: **1)** We study the robustness of an existing retrieval-augmented captioning model SmallCap and find it is not robust to processing randomly retrieved content. **2)** We identify that tokens that frequently occur in the retrieved captions, i.e. majority tokens, have high attribution scores with regard to the tokens generated by the model. This phenomenon suggests heightened sensitivity and copying. **3)** Training with sampled retrieved captions from a larger list instead of with fixed top-k relevant captions improves model robustness, yielding better generalization and out-of-domain performance.[1]

5.2 Related Work

Robustness of retrieval-augmented models. Retrieval-augmented generation (RAG) involves enhancing the generation process by incorporating retrieved information from an external datastore as additional context to the input (Lewis et al., 2020). RAG models have shown to improve performance across a variety of NLP tasks (Mialon et al., 2023). However, RAG models can overly rely on retrieved information, resulting in inaccurate generation when the retrieved context is flawed (Yan et al., 2024; Yoran et al., 2023).

Recent efforts aim to enhance RAG model robustness against misguided or hallucinated generations. One approach involves filtering retrieved content (Wang et al., 2023b; Yoran et al., 2023; Yasunaga et al., 2023; Yan et al., 2024; Asai et al., 2023) by applying or training an additional evaluator. Another direction focuses on improving robustness during the training of the generation model itself. Specifically, for retrieval-augmented question answering with large language models, Yoran et al. (2023) propose continued training with a synthetic dataset that contains both relevant and irrelevant context, while Cuconasu et al. (2024) suggests incorporating irrelevant documents. In retrieval-augmented translation, robustness is improved through shuffling retrieved translations (Hoang et al., 2022), ensemble model decoding (Hao et al., 2023), and controlled interactions between source and retrieved translations (Hoang et al., 2023).

Retrieval-augmented image captioning. Image captioning is the task that describes the visual contents of an image in natural language (Xu et al., 2015; Osman et al., 2023). Recent studies have integrated RAG into this field. Sarto et al. (2022) and Zhou and Long (2023) experimented with retrieving similar or style-aware images before generating captions. Li et al. (2023a) introduced a lightweight image captioning model that utilizes

[1]We release the code at https://github.com/lyan62/RobustCap

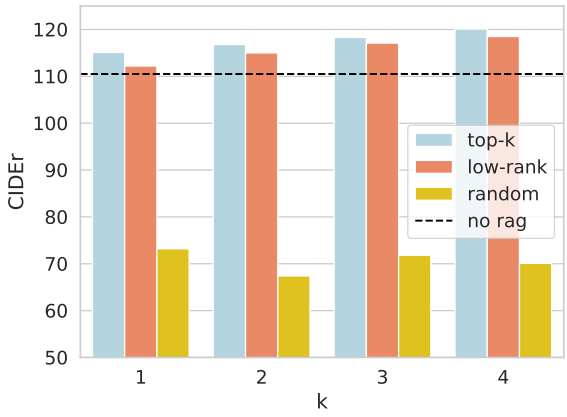


Figure 5.2: CIDEr evaluation of SmallCap on the COCO validation set using the top-k, low(er)-ranked, randomly retrieved captions, against a baseline without retrieval augmentation. Performance drops by up to 50% when using randomly retrieved captions compared to baseline, suggesting that the model is not robust.

retrieved concepts. More related to our work, Ramos et al. (2023a) developed end-to-end encoder-decoder models that attend to both the image and retrieved caption embeddings.

In particular, the SmallCap model (Ramos et al., 2023c), presenting retrieval augmentation in image captioning could reduce trainable parameters and adapt to out-of-domain settings. The model utilizes frozen unimodal models, incorporating a pre-trained encoder and decoder connected by trainable cross-attention layer.

However, it still remains unclear how retrieved captions influence the generation of captions in retrieval-augmented image captioning, especially concerning visual inputs. Additionally, the evaluation and enhancement of the robustness of these models are still under-explored.

5.3 Robustness of Retrieval-Augmented Image Captioning

To evaluate the robustness of the SmallCap retrieval-augmented caption model (Ramos et al., 2023c), we conduct controlled experiments and observe its resilience to changes in (1) the order of the retrieved captions and (2) the content relevance of the retrieved captions.

5.3.1 Robustness Evaluation

For a given image, SmallCap is augmented with a sequence of k retrieved captions that are combined into an input for the language model decoder: *"Similar images show* $cap_1, cap_2, ..., cap_k$*. This image shows ..."*. The retrieved captions are obtained through image-to-text retrieval using CLIP embeddings (Radford et al., 2021), and are sorted according to their relevance, i.e., cosine similarity. From the sorted retrieved captions, we retain the most similar captions as the retrieval list. In this regard, the top-k retrieved candidates are the first k captions in the list, and the low-ranked captions are the last-k captions in the list. SmallCap uses the top-k retrieved captions in the prompt by default.

Context order. When prompting the model to generate a caption for a given image, we can change the order of the retrieved captions by **permuting** or **reversing** them. We evaluate the effect of the order changes in two settings: one with a model trained using the top-k retrieved captions (default), and another that is also trained with permuted or reversed retrieved captions.[2]

Content relevance. To evaluate how robust the model is towards noise in the retrieved captions, we are curious to see how the model performs when (1) captions are randomly retrieved, i.e. likely to be irrelevant for the given image (2) only low-ranked retrieved captions are available. Here the randomly retrieved captions are those retrieved with another image. For low-ranked captions, we take the lowest-ranked k captions from the retrieval list that consists of top seven relevant captions.

5.3.2 Experimental Setup

In the experiments, we set $k = 4$ as it has been demonstrated as the optimal number of captions by Ramos et al. (2023c). We evaluate SmallCap models with both OPT-350M (Zhang et al., 2022b) and GPT-2 (Radford et al., 2019) as the decoder models. For the image encoder, we use ResNet-50x64 (He et al., 2016) and CLIP-ViT-B/32 (Radford et al., 2021) as the retrieval encoder. We keep the same model setting in the following sections unless stated otherwise.

Data and metrics We first evaluate the robustness of Smallcap on COCO validation set for *in*-domain evaluation. Then we evaluate on NoCaps (Agrawal et al., 2019), which contains *In, Near* and *Out*-of-domain data, and serves as a challenging dataset designed

[2]For the model trained with default order—top four captions, we use the pretrained checkpoints from HuggingFace: https://huggingface.co/Yova/SmallCap7M, https://huggingface.co/Yova/SmallCapOPT7M

Retrieval Order		LM Backbone	
Train	Eval	GPT-2	OPT
default	default	116.4	120.3
	permute	116.2	120.1
	reverse	115.8	119.7
permute	permute	**117.2**	**120.4**
reverse	reverse	116.4	**120.7**

Table 5.1: CIDEr evaluation on the COCO validation set with GPT-2 and OPT variants of SmallCap when manipulating the order of the top-k retrieved captions.

to assess the generalization capabilities of models trained on COCO. For both datasets we use the validation set experimenting with different number of retrieved captions, i.e. different k values. We report peformance using CIDEr score (Vedantam et al., 2015).

5.3.3 Order Robust but Content Sensitive

Order robust. From the results in Table 5.1 and Table 5.2, we observe that SmallCap is indeed robust to the order of the retrieved texts. Permuting the order of the captions during training and evaluation show 1 CIDEr point improvement for COCO (Lin et al., 2014) and $2-3$ CIDEr score increase for NoCaps (Agrawal et al., 2019). This indicates that if multiple captions are used for augmentation, then permuting their order helps.

Content sensitive. Figure 5.2 shows that when using randomly retrieved captions instead of the top-k most relevant captions, performance drops drastically compared to the no-retrieval baseline.[3] This implies that smallcap lacks resilience to noise in the retrieved captions, and the irrelevant context has the potential to mislead the model, resulting in inaccurate predictions. When prompting with low-ranked retrieved captions, while performance slightly decreases, the retrieval-augmented model still outperforms the one without retrieval.

5.4 Majority Tokens Explain Behavior

To better understand how each token of the retrieved content relates to the observed sensitivity discussed in the previous section, we hypothesize that the model is driven

[3]Here the top and low ranked captions are obtained from a list of top-seven captions retrieved captions ordered by their cosine similarity to the image embedding.

Retrieval Order		LM Backbone					
		GPT-2			OPT		
Train	Eval	In	Near	Out	In	Near	Out
default	default	80.1	79.4	69.6	91.0	84.4	76.3
	permute	**81.6**	**79.8**	68.5	**92.5**	**84.5**	75.8
	reverse	80.2	79.3	68.4	**92.0**	84.4	**76.6**
permute	permute	**81.5**	79.7	**69.8**	**94.2**	84.0	**79.4**
reverse	reverse	**80.4**	**80.1**	68.4	**92.5**	**85.6**	75.9

Table 5.2: Evaluation on NoCaps using CIDEr score with the GPT-2 and OPT variants of SmallCap when manipulating the order of the top-k retrieved captions.

by the presence of majority tokens. In other words, when the model is prompted with retrieved captions, we assume that the predicted tokens are influenced by the tokens that appear in the majority of the retrieved captions. To test this assumption, we propose a majority voting analysis, followed by input attribution, and an attention analysis of the model behavior.

5.4.1 Majority Tokens

We first introduce the definition of majority tokens. Let $R = [T_1, \cdots, T_n]$ represent a retrieved caption R, which contains a sequence of n tokens. For a given image, we assume that a total of K retrieved captions are used in the model prompt: R_1, R_2, ..., R_K. For each token T_i in the set of unique tokens from the retrieved captions, we define T_i as a *majority token* (denoted as T_M) if T_i appears in more than half of the retrieved captions[4], i.e., $C_{T_i} > \frac{K}{2}$ where C_{T_i} is the number of retrieved captions that contains token T_i as in Equation 5.1:

$$C_{T_i} = \sum_{l=1}^{K} \mathbb{1}[T_i \in R_k] \tag{5.1}$$

For a generated caption $Y = [y_1, \cdots, y_n]$ in the evaluation data, we can calculate the majority-vote probability $P_{T_M \in Y}$ as the probability of the majority token T_M appearing in the generated caption.

[4]Note that we remove the stop words in the retrieved captions when determining the majority tokens. The stop words are filtered from the top-100 most frequent tokens in the COCO dataset, where we manually remove meaningful tokens such as "man", "two" from the list. Please see the Appendix 5.8.1 for the complete list.

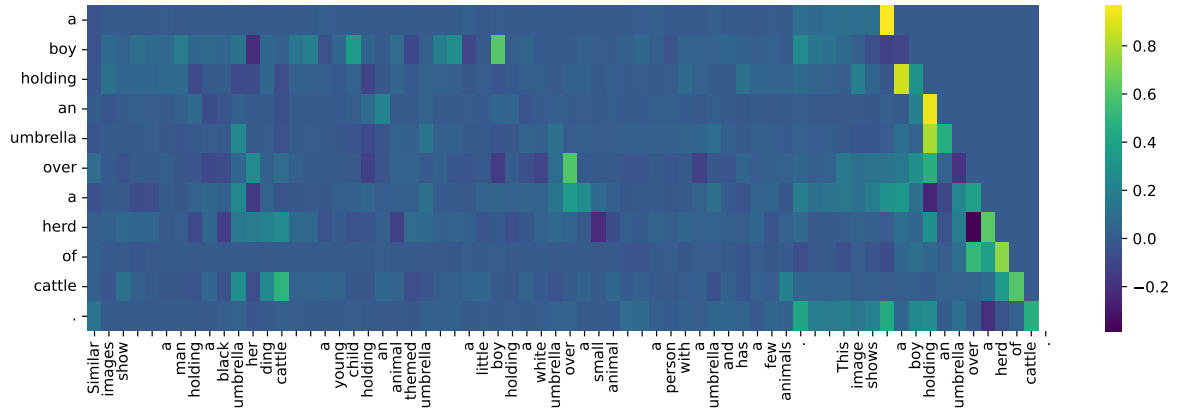


Figure 5.3: Input attribution for each generated token (y-axis). The brighter the color, the more greater the attribution from the input token. We observe high attribution scores to "umbrella", "boy", "cattle", and "over".

We expect that the higher the value of $P_{T_M \in Y}$, the more likely it is that the model is generating captions based on the majority tokens.

5.4.2 Experimental Setup

We test our majority vote assumption with a controlled experiment. Specifically, we analyze the predictions of the model in two settings, each provided with $K = 3$ retrieved captions to ensure the presence of a majority token:

2 Good 1 Bad (2G1B): The retrieval set contains two relevant captions and one irrelevant caption;

2 Bad 1 Good (2B1G): The retrieval set contains two irrelevant captions and one relevant caption.

The assumption is that, if there is a majority voting behavior with respect to the retrieved captions, the model will copy such majority tokens to the final output. The distinction will be clear in this setting — in the setup 2B1G, if the model is robust to the retrieved context, the model will focus more on the good caption instead of the majority tokens in the two bad captions.

We use the COCO evaluation set and the pretrained checkpoint with the OPT decoder of Ramos et al. (2023c) for this analysis. Good captions are obtained using the top-two and top-one retrieved captions, respectively, for a given image. Bad captions are obtained by retrieving one or two captions, respectively, from a randomly selected image.

Results. We find that the probability of majority vote in the 2G1B setting is 86.47%. This high probability suggests that the majority tokens in the good captions could be being used to guide the model generation. In the 2B1G setting, the model is much less likely to generate majority tokens from the bad captions, indicating some robustness in not always following them. However, 20.84% of the time, the model can still be misled by their appearance, resulting in the majority tokens being copied into the model output.

5.4.3 Input Attribution with Integrated Gradients

To better understand the role of majority tokens in model generation, we use integrated gradients (Sundararajan et al., 2017) for input attribution analysis. This enables us to examine the influence of each individual token in the retrieved captions on the model prediction.

Attribution visualization. Figure 5.3 shows an example of an attribution visualization, where the attribution score of each input token (x-axis) is computed at each generation step (y-axis). Bright color cells correspond to high attribution to the input token. High attribution scores to the same tokens seen in the retrieved captions may indicate copying. Negative attribution scores are observed at contradicting tokens observed in the retrieved captions to the current generation. Negative scores are observed at token "her" when model is predicting the token "boy" and at token "small" when predicting "herd". Additional input attribution visualizations can be found in Appendix 5.8.2.

Quantitative analysis. We also quantitatively analyze the impact of majority tokens by calculating pairwise attribution scores between tokens in retrieved captions and those predicted by the model. Higher attribution values suggest greater sensitivity to the input token (Ancona et al., 2018). Figure 5.4 shows the distribution of the pairwise attribution scores for the 2B1G setup. It is clear that the model is sensitive to the majority tokens, especially when the generated token exists in the retrieved captions. Such behavior indicates weak robustness: we would not expect a robust model to be distracted by the tokens from the two irrelevant retrieved sentences at inference time. To better visualize the impact, we show distribution of original attribution values and the absolute values (Ancona et al., 2018) across all evaluation samples.

5.4.4 Attention and Model Behavior

Finally, we visualize the self-attention and cross-attention to locate the heads and layers in the SmallCap-OPT125M model that may contribute to the majority voting behaviour when generating a caption. This is crucial because all interactions between captions (self-attention) and images (cross-attention) take place in this stage.

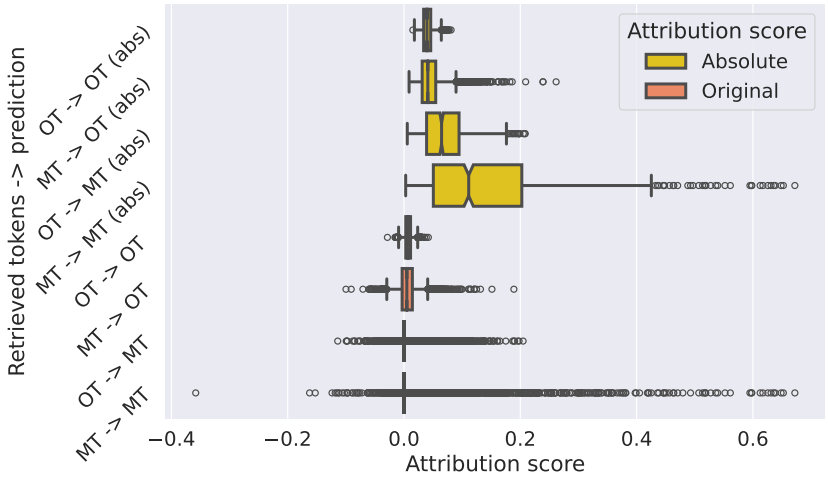


Figure 5.4: Pairwise average attribution score between retrieved and generated tokens in the 2B1G setup. MT: majority tokens in the retrieved captions. OT: all other tokens. The larger pairwise attribution values shows that the majority tokens have a larger impact during generation than the other tokens in the retrieved captions.

Distribution of max attention occurrence. We partition the text input prompt into five distinct segments: begin of the sentence token (<BOS>), prompt tokens before retrieved k captions (*prefix*), i.e. "Similar image shows", the retrieved captions (*retrieval*) $cap_1, \cdots cap_k$, prompt tokens before generation (*suffix*), i.e. "This image show", and the *generation* itself. For image patches, we segment them into two pieces – the CLS output embedding, and the set of patch output embeddings.

Let S_n denote the sets of indices, where $n = 1, 2, \ldots, 5$ for five segments. For the text input, each segment S_n contains the indices of the tokens in each segment. To track the occurrence of max attention values in S_n, we define the indicator function $\mathbb{1}[I_n(i,j)]$ as follows:

$$\mathbb{1}[I_n(i,j)] = \begin{cases} 1 & \text{if } \arg\max_z Att(j,z)_i \in S_n \\ 0 & \text{otherwise} \end{cases}, \tag{5.2}$$

where $\arg\max_z Att(j,z)_i$ is the index of the input with the maximum attention score for sample i.

For self-attention between the textual tokens, $Att(j,z)$ represents the attention score between the j^{th} generated token and the z^{th} text context token, denoted as $SA_{text}(j,z)$. For cross-attention between the decoder and the image representations, we report both a text-centric and an image-centric analysis. The text-centric analysis $XA_{text}(j,z)$ measures the attention between the j^{th} image patch and the z^{th} text token, to identify which segments of the text have the highest cross-attention scores in relation to the image. In the

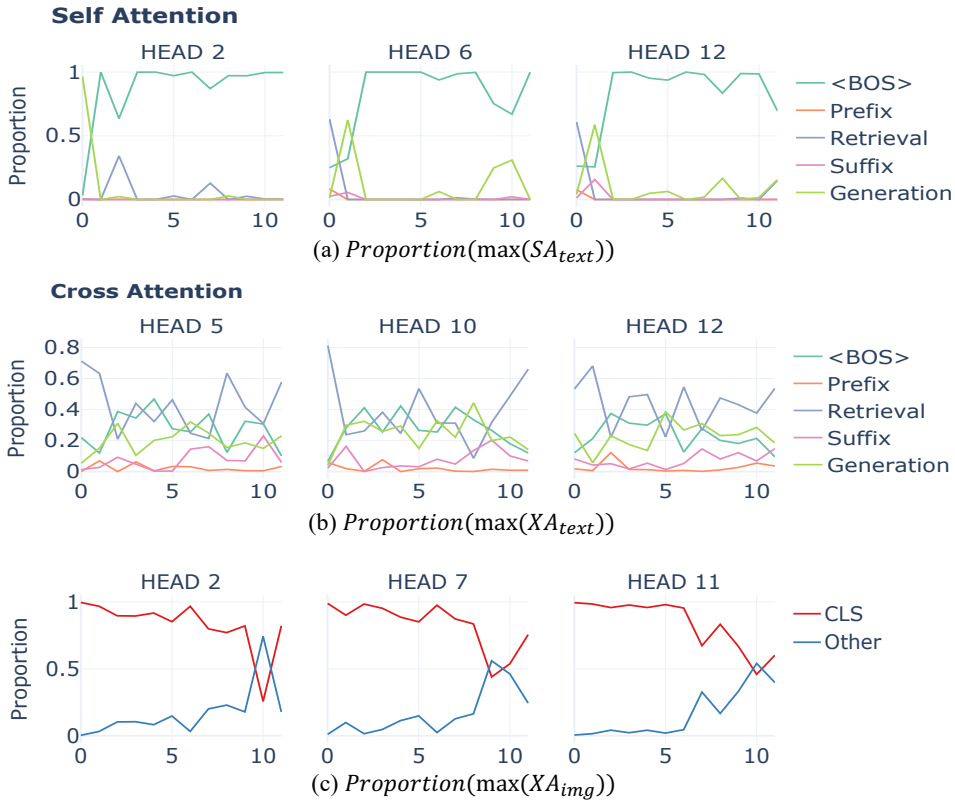


Figure 5.5: Statistics of all maximum attention scores' distribution across different layers and heads from self and cross attention. XA denotes cross attention, while SA signifies self-attention. img represents the distribution of maximum attention scores across image patches, whereas $text$ pertains to the distribution of maximum attention scores across text tokens.

image-centric analysis $XA_{img}(j, z)$, we measure the attention between the j^{th} generated token and the z^{th} image patch. We now redefine the S_n notation to let S_1 represent the CLS output embedding, and S_2 represent the set of image patch embeddings, respectively. This allows $XA_{img}(j, z)$ to identify if the CLS patch embedding receives the highest cross-attention scores in relation to the generated tokens, or if it is the actual image patch embeddings. For each analysis, $SA_{text}(j, z)$, $XA_{text}(j, z)$, and $XA_{img}(j, z)$, we calculate the proportion of occurrences of the maximum score in S_n by averaging through all generated tokens for a dataset.

Self-attention. We gather attention scores between the generated tokens and context tokens, and categorize the distribution of the maximum scores into the five text segments

(BOS, prefix, retrieval, suffix, and generation).

Figure 5.5(a) illustrates the changes in the distribution of maximum self-attention scores in each layer of the decoder language model. Notably, at the initial layers, a majority of attention heads exhibit heightened focus on retrieved captions or the current context for generation. However, after the second layer, we observe an increased emphasis on the beginning of sentence token (</s>). This behavior is consistent with prior research on the attention mechanism of GPT-2 (Vig and Belinkov, 2019). Figures 5.9 and 5.11 show the behaviour for all self-attention heads in for the GPT and OPT model variants, respectively.

Cross-attention. Similar to the self-attention behaviour, we categorize the occurrence of the maximum cross-attention to the five text segments. As shown in Figure 5.5(b), in most attention heads, the cross-attention attains its maximum value between the image and the retrieved captions or between the image and the generated tokens. Figures 5.10a and 5.12a show the text-centric analysis for all cross-attention heads for the GPT and OPT backbones.

Finally, we inspect whether the model focuses on the CLS patch or actual image patches. In Figure 5.5(c), we observe that the model only pays maximum attention to the image patches in the final layers (the blue line). Figures 5.10b and 5.12b show the full results for the image-centric analysis.

Overall, these observations show that the model attends to both modalities during the caption generation process. However the lack of strong cross-attention to actual image patches suggests that the model is misled by text prompts, even when irrelevant information is absent in the provided image.

5.5 Improving Robustness to Retrieval via Sampling

In order to improve the robustness of the model to potentially noisy captions, we propose to randomly sample the captions from a larger retrieval list for a given image, instead of training with only the top-k retrieved captions. In this manner, the model can learn from more diverse context that includes both top- and lower-ranked captions.

5.5.1 Experimental Setup

Inspired by Hoang et al. (2022), we experiment with two sampling methods during training to improve retrieval robustness.

Sample-k training. We sample k captions randomly from the top-N=7 retrieved captions during training[5]. Following Ramos et al. (2023c), we train SMALLCAP with the OPT-350M decoder on the COCO captioning dataset (Chen et al., 2015) for 10 epochs on a NVIDIA A100 40GB GPU with the default learning rate of 1e-4 and batch size of 64. We experiment with k in the range of 1–4.

Controlled sample-k training (c-sample-k). Aiming to train the model that better distinguishes irrelevant context, we design a controlled sampling process — selecting $k-1$ randomly from the larger list while keeping the top relevant caption of the image during training. We train the model with same hyperparameters and dataset as sample-k.

5.5.2 Evaluation and Results

In addition to the COCO and NoCaps validation set, we evaluate the *Out*-domain performance of the model using VizWiz caption dataset (Gurari et al., 2020) and report CIDEr scores.

		COCO Eval		
Model	k	top-k	last-k	random
top-k	1	115.1	112.2	73.2
sample-k	1	**116.0**	**115.0**	**98.9**
top-k	2	116.8	115.0	67.4
sample-k	2	**117.4**	**116.8**	**84.6**
top-k	3	118.3	117.1	71.8
sample-k	3	**118.5**	**117.3**	**77.6**
top-k	4	120.1	117.1	70.1
sample-k	4	119.2	**118.6**	**73.1**
c-sample-k	4	119.3	**118.9**	**72.6**

Table 5.3: CIDEr scores when training on the top-k, sample-k and c-sample-k captions. Training by sampling the retrieved captions almost always outperforms SMALLCAP for all k values. It also reduces the gap between using top-relevant and low-ranked retrieved captions. Results are averaged over three seeds. Improved scores are in **bold**.

[5]We sample from the top-N=7 for alignment with the baseline; see the Appendix for an ablation on varying N.

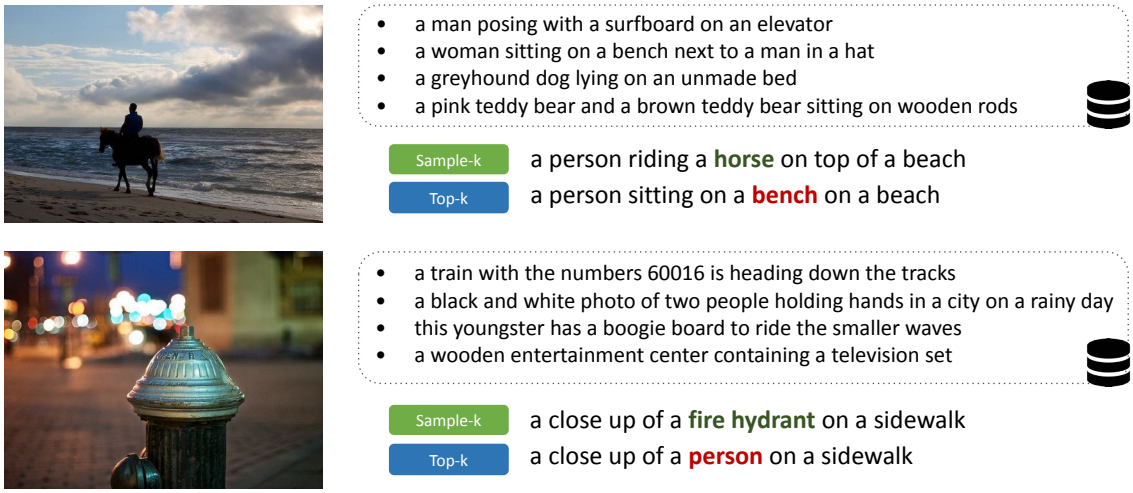


Figure 5.6: Qualitative examples of generated captions when **randomly** retrieving four captions for a given image using a model trained with either the Sample-k or the Top-k method.

Sample-k training improves model robustness to random retrieved captions. As shown in Table 5.3, incorporating sampled retrieved captions into training consistently enhances performance across various k values. The improvement is particularly notable when captions are randomly retrieved, suggesting the model is now better able to ignore irrelevant context. If we compare across different values of k, sampling mitigates the model's sensitivity to the number of retrieved captions, outperforming top-k training. For instance, it achieves comparable performance with a smaller k value than in the case of top-k training. Furthermore, the gap between using the top-k vs. the last-k retrieved captions is reduced with sample-k training: the maximum gap is reduced from 3.0 to 1.0 CIDEr points, indicating increased model robustness, even with lower-ranked retrieved captions. Figure 5.6 and 5.14 show qualitative examples of the improved robustness to randomly retrieved examples.

Sampling improves cross-domain evaluation. We also evaluate on VizWiz and NoCaps to measure cross-domain performance (Table 5.4). This is a more realistic setting where retrieved captions are out-of-domain and could be more noisy and less relevant. The application of sampling improves across all values of k for Vizwiz. On the NoCaps dataset, with the COCO datastore, sampling consistently improves near and out-domain performance, suggesting increased robustness to noisy retrieval context. This is consistent with the benefits of sampled training demonstrated in cross-domain machine translation by Hoang et al. (2022). If we use a larger datastore that incorporates internet-derived

		VizWiz	NoCaps			NoCaps (+Web)		
Model	k		In	Near	Out	In	Near	Out
top-k	1	31.3	85.0	74.3	62.3	84.1	80.7	81.5
sample-k	1	**32.3**	**87.0**	**75.7**	**63.6**	**87.8**	**81.2**	77.5
top-k	2	33.7	85.0	74.3	62.3	90.5	86.2	89.5
sample-k	2	**34.0**	**87.8**	**77.4**	**67.6**	**90.6**	85.3	86.7
top-k	3	35.0	87.4	79.6	68.3	91.7	88.3	89.9
sample-k	3	**35.4**	**88.7**	**80.3**	**69.4**	**92.6**	88.0	**90.0**
top-k	4	35.5	87.4	79.6	68.3	94.2	89.4	91.2
sample-k	4	**35.7**	**89.7**	**80.9**	**71.1**	**94.8**	**89.5**	**93.1**
c-sample-k	4	**36.0**	**90.1**	**81.3**	**71.5**	**94.5**	**90.0**	**93.3**

Table 5.4: Training with sampled retrieval always outperforms top-k retrieval for all values of k on the out-of-domain VizWiz and NoCaps datasets. The gains are smaller when using a larger datastore (+Web) but it still improves out-domain performance when retrieving more captions. Improved scores are in **bold**.

captions (+Web), this consistently improves in-domain performance. Retrieval constraints are alleviated for near and out-domain samples with the larger datastore, where we see smaller gains with sample-k. See qualitative examples in Figure 5.13 in Appendix 5.8.3.

Controlled sampling further improves cross-domain evaluation. Finally, on top of our best performing sample-k model, controlled sample-k further improves performance for both NoCaps and VizWiz. This suggests that incorporating both top-relevant and low-ranked captions during training aids the model in distinguishing irrelevant context.

5.6 Discussion

Majority tokens are reliable hints during training. To better understand why the model relies on majority tokens during generation, we calculate the probability that majority tokens in the retrieved captions overlap with the ground truth captions ($T_M \in GT$), and with the predicted tokens ($T_M \in Pred$). Table 5.5 shows that in 88%–99% of the training examples, the majority tokens in the retrieved captions are also present in the ground truth captions. This suggests that the model can develop a bias towards majority tokens due to the fact that they are so often present in the ground truth during training. This analysis also clarifies the decrease in the model's robustness as k increases

	k=2	3	4
$T_M \in GT_{train}$	88.0	97.5	99.2
$T_M \in GT_{val}$	74.7	86.5	91.0
$T_M \in$ Pred	82.8	93.4	96.7
$T_M \in$ Pred (sample-k)	81.9	93.3	96.6

Table 5.5: Percentage of samples in the COCO train and validation set where the majority token of the retrieved captions are present in the ground truth compared to the percentage of their presence in prediction.

when randomly retrieving captions. This is because a higher k only adds noise without providing useful majority tokens. The use of sampling during training exposes the model to more diverse context, which leads to a slightly increased level of selectivity.

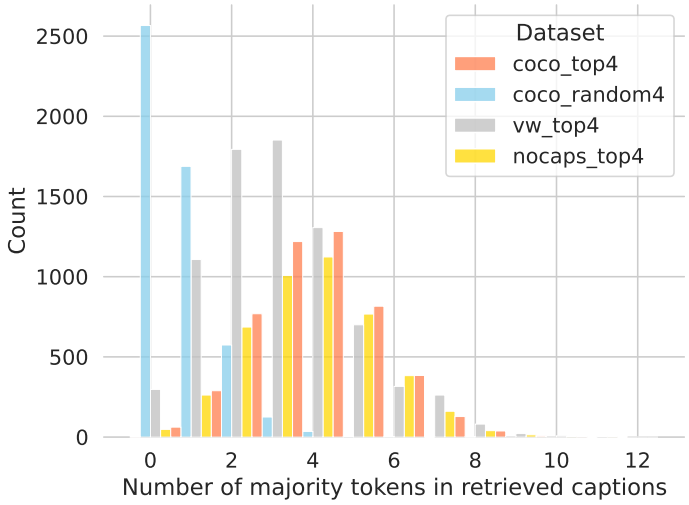


Figure 5.7: Distribution of number of majority tokens in the retrieved captions for the COCO, VizWis, and NoCaps evaluation datasets. For the COCO dataset, we also show the difference between retrieving the top-4 captions against four randomly selected captions.

In Figure 5.7, we show the variation in the distribution of majority tokens across various evaluation datasets. When captions are randomly selected for the COCO evaluation data, there are fewer majority tokens in the retrieved captions. This presents a challenge for the model in making use of the retrieved captions, which accounts for the performance decrease shown in Figure 5.1. For evaluation, with the same value of k, the fewer the number of majority tokens in the retrieved captions, the harder it is for the model to "copy" those tokens to the final output. In such scenarios, we obtain bigger improvements

with the sample-k training.

5.7 Conclusion and Future Work

We studied the robustness of the state-of-the-art retrieval-augmented image captioning model SmallCap and provide an through analysis and explanation of how retrieved captions effect the final prediction. Our exploration shows that SmallCap is robust to the order of the retrieved captions, but it is sensitive to retrieval noise, which has implications for using retrieval-augmented models in new domains. With extensive input attribution analysis, we show that such sensitivity is due to majority tokens in the retrieved captions. We demonstrate a more retrieval robust model can be trained with sampling methods during training. We expect that our analysis can inspire better retrieval-robust captioning models in the field.

In the future, we will investigate whether the majority voting behaviour is exploited in other retrieval-augmented captioning models. We hope to further explore if other techniques such as token-dropping or prefix-tuning would further improve retrieval robustness.

Ethics Statement

We acknowledge the potential risks of hallucination and biases introduced by retrieval augmentation in captioning models. Misleading tokens from the retrieved captions could cause the model to generate captions describing nonexistent entities or objects in images (Liu et al., 2024; Rohrbach et al., 2018). This could have adverse effects, such as propagating systematic biases present in the datastore used for retrieval (Foulds et al., 2024).

Despite the exploration in our work, we acknowledge that no system is perfect, and undesirable biases may still be present with our methods. We emphasize the need for continued research into techniques for identifying and mitigating hallucination and bias in retrieval-augmented models (Foulds et al., 2024; Deng et al., 2024). We also stress the importance of responsible deployment, with human oversight and content moderation pipelines.

As researchers, we have an ethical obligation to be transparent about the potential risks and limitations of our work. We welcome further scrutiny and discussion around these critical issues within the research community.

Limitations

We evaluate the robustness of a single retrieval-augmented image captioning model in this study. Given variations in training process and model structures, the observed model

behavior may be specific to our chosen model. Applying the same analysis to other models would be useful for a deeper understanding regarding explainability and interpretation of retrieval augmented image captioning models, which we leave for future work.

For all experiments in our study, we employ the same CLIP-ViT-B/32 backbone as the image encoder. Investigating how model robustness varies with different visual encoders would enhance the scope of our study.

While training with sampling improves model robustness, it is intuitive that introducing more noise during training makes the task more challenging. In all our experiments, we train the model for same number of epochs as SmallCap, therefore it is not clear if the model would gain more robustness if trained longer. We are curious if there exists an optimal balance between training time and the level of noise exposure for achieving model robustness.

Acknowledgments

We thank Lei Li and the CoAStal and LAMP groups for feedback. Wenyan Li is supported by the Lundbeck Foundation (BrainDrugs grant: R279-2018-1145) and a research grant (VIL53122) from VILLUM FONDEN. Jiaang Li is supported by Carlsberg Research Foundation (grant: CF221432). Rita Ramos is supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (i.e., the Center For Responsible AI), and also by Fundação para a Ciência e Tecnologia (FCT), through the project with reference UIDB/50021/2020 (DOI:10.54499/UIDB/50021/2020) and the Ph.D. scholarship with reference 2020.06106.BD.

5.8 Appendix

5.8.1 Majority Tokens

Stop words list In this section, we present the stop words that were filtered from the COCO dataset in the experiments described in Section 5.4.2:

['out', 'some', 'of', 'is', 'while', 'are', 'with', 'down', 'has', 'over', 'the', 'next', 'up', 'near', 'several', 'other', 'at', 'top', 'from', 'in', 'on', 'a', 'there', 'an', 'to', 'and', 'her', 'front', 'by', 'for', 'his', 'it']

5.8.2 More Visualization

Input Attribution with Integrated Gradients

In Figure 5.8, we show more attribution visualization for the experiment setup 2B1G in Section 5.4 where high attribution scores are observed in the majority tokens and mislead the model to generate incorrect captions.

Attention

In Figure 5.9, Figure 5.10, Figure 5.11 and Figure 5.12 , we depict the distributions of both self-attention and cross-attention scores across various heads and layers for SmallCap with GPT-2 and OPT decoder variants.

5.8.3 Qualitative examples

We show more qualitative examples in Figure 5.13 and Figure 5.14.

5.8.4 More results

Order robustness evaluation In Table 5.6 and Table 5.7, we provide both CIDEr and BLEU4 scores for order robustness evaluation (Section 5.3).

Retrieval Order		SmallCap LM	
Training	Evaluation	GPT-2	OPT
default	default	116.4/36.1	120.3/37.1
	permute	116.2/36.0	120.1/37.0
	reverse	115.8/36.0	119.7/36.8
permute	permute	**117.2/36.4**	**120.4/37.2**
reverse	reverse	116.4/36.1	**120.7**/37.0

Table 5.6: Results of manipulating the order of the top-k retrieved captions by either randomly permuting or reversing the list. We report CIDEr/BLEU4 scores on the COCO validation set using either a GPT-2 or OPT backbone in the SmallCap model.

Number of retrieved captions for sample-k training We experiment with different size of the retrieval candidate list from which we randomly select captions for sample-k training (Table 5.8).

Model	Retrieval Order	In	Near	Out
GPT2	default	80.1/37.9	79.4/35.9	69.6/25.3
	permute	**81.5/38.8**	79.7/36.6	**69.8/26.2**
	reverse	80.4/38.4	**80.1/36.3**	68.4/25.1
OPT	default	91.0/27.1	84.4/23.8	76.3/15.0
	permute	**94.2/28.6**	84.0/25.0	**79.4/15.8**
	reverse	92.5/28.4	**85.6/25.3**	75.9/14.2

Table 5.7: Complete results with both CIDEr/BLEU4 on the NoCaps dataset when evaluated with different order of the top-four retrieved captions. The order applies to both train and evaluation stage.

			COCO		
Size	k	VizWiz	top-k	last-k	random
7	4	36.0	119.2	117.1	71.0
10	4	36.0	**119.3**	**118.3**	67.6
50	4	33.9	118.1	117.7	**81.2**

Table 5.8: CIDEr score when sampling from different size of retrieval candidates. We see more improvements on random k evaluation while almost keeping the same level of in-domain performance. With more noise involved during training, we would expect a longer training time would yield more robust performance.

Percentage of tokens that are likely to be copied In Table 5.9 we show the percentage of tokens that are likely to be copied from retrieved captions averaging through all samples in the validation set. Majority tokens takes more than half of the copied tokens.

Comparison with other methods Inspired by Yoran et al. (2023), we have considered intentionally including less relevant captions by including one irrelevant caption, one low-ranked caption, and top-2 relevant captions instead of using top-4 retrieved captions. However, in our preliminary experiments, this strategy does not perform as well as the sampling approach, likely due to the high noise level it introduced.

	k=1	2	3	4
$T_R \in$ Pred	49.1	63.3	69.8	75.7
$T_R \in$ Pred (sample-k)	46.0	61.5	69.5	74.0
$T_M \in$ Pred	-	33.1	45.7	54.5
$T_M \in$ Pred (sample-k)	-	32.5	45.3	53.3

Table 5.9: Percentage of tokens in the predicted caption that are likely copied from majority tokens in retrieved captions in the COCO validation set. T_R represent tokens in retrieved captions. T_M represent the majority tokens in retrieved captions.

	COCO Evaluation			NoCaps Evaluation		
Method	top-k	last-k	random	In	Near	Out
top-4	120.1	117.1	70.1	87.4	79.6	68.3
sample-4	119.2	118.6	73.1	89.7	80.9	71.1
mixed-4	119.2	118.1	66.7	59.9	57.9	39.4

Table 5.10: CIDEr on COCO and NoCaps.

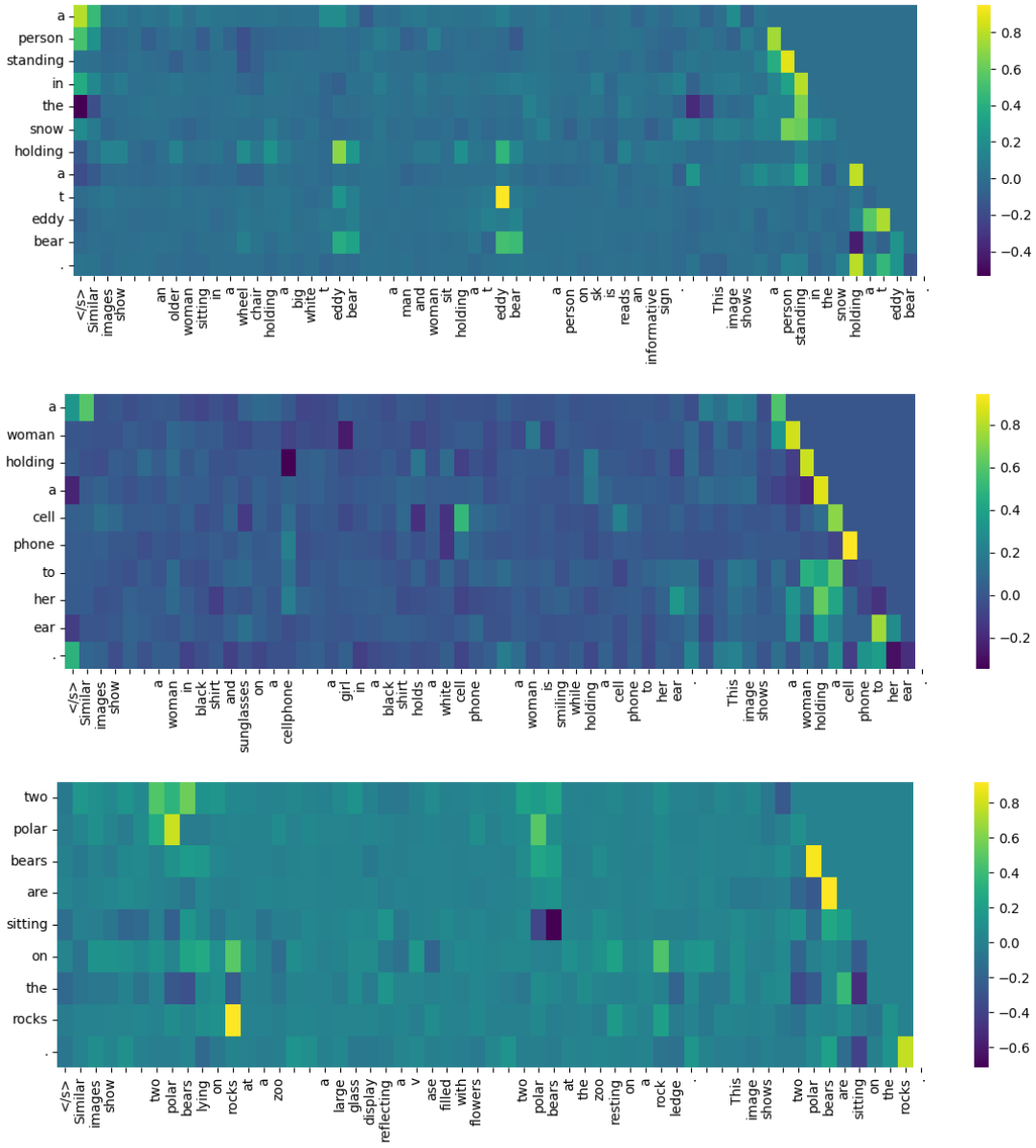


Figure 5.8: Attribution visualization with few more examples. Here the model prediction is misled by the majority tokens in the 2B1G setting.

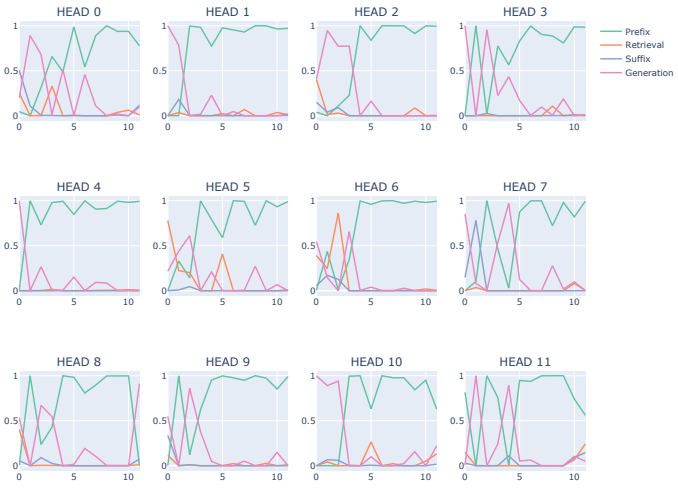
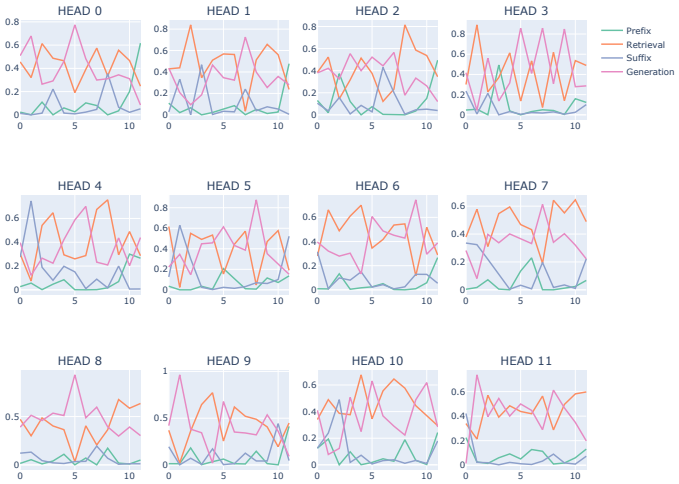
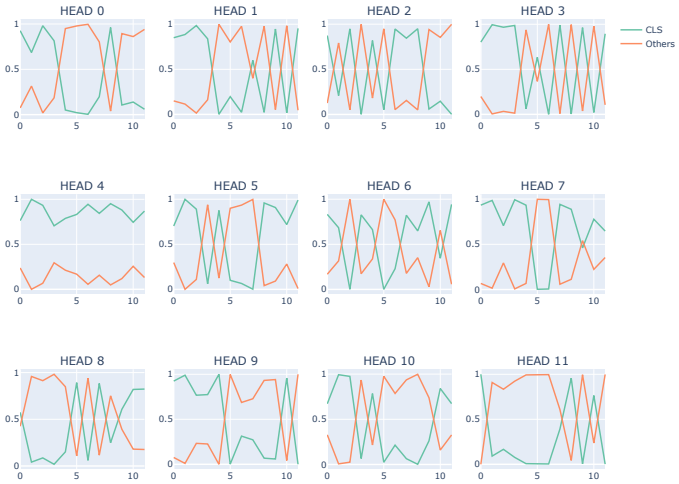


Figure 5.9: Self attention distribution in SmallCap (GPT2 variant). Statistics of max attention scores from different layers and heads, showing the proportion of attention scores belonging to different parts.



(a) Distribution of max attention scores of the interaction between various parts of text prompt and image patches.



(b) Distribution of max attention scores of the interaction between two types of image patches (cls, others) and all text tokens.

Figure 5.10: Cross attention distribution in SmallCap (GPT2 variant). Statistics of max cross-attention scores from different layers and heads, showing the proportion of attention scores belonging to different parts of the multimodal input.

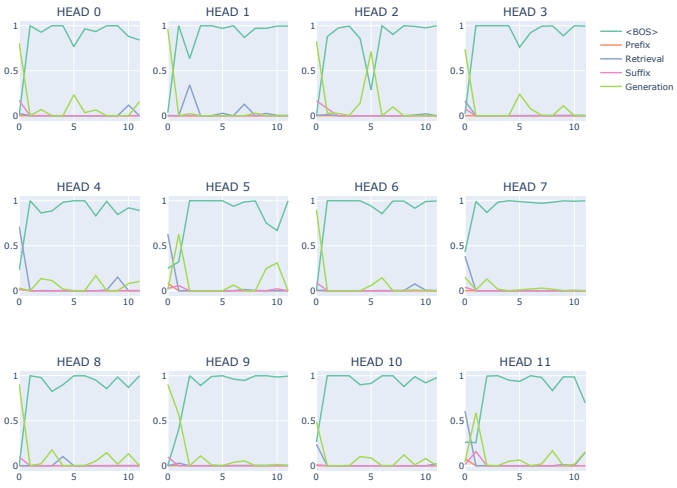
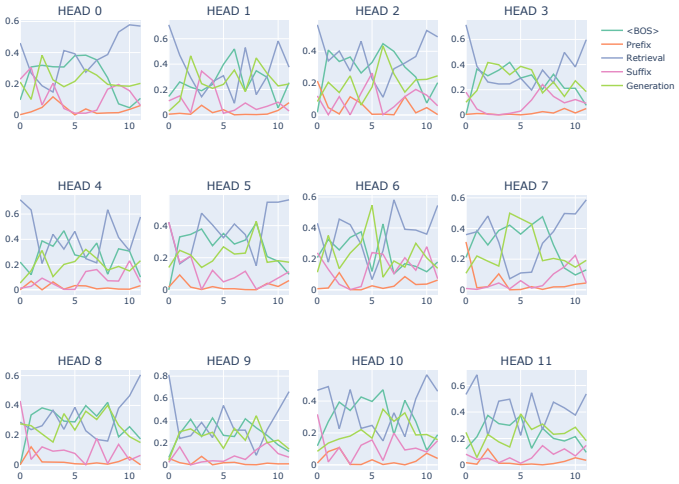
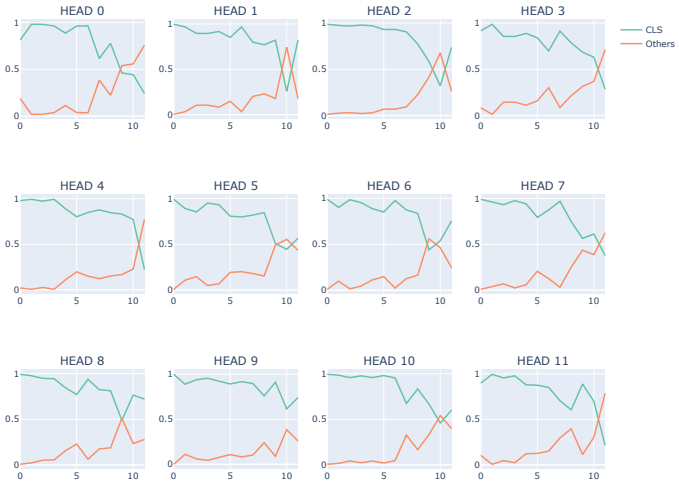


Figure 5.11: Self attention distribution in SmallCap (OPT-125M variant). Statistics of max attention scores from different layers and heads, showing the proportion of attention scores belonging to different parts.



(a) Distribution of max attention scores of the interaction between various parts of text prompt and image patches.



(b) Distribution of max attention scores of the interaction between two types of image patches (cls, others) and all text tokens.

Figure 5.12: Cross attention distribution in SmallCap (OPT-125M variant). Statistics of max cross-attention scores from different layers and heads, showing the proportion of attention scores belonging to different parts of the multimodal input.

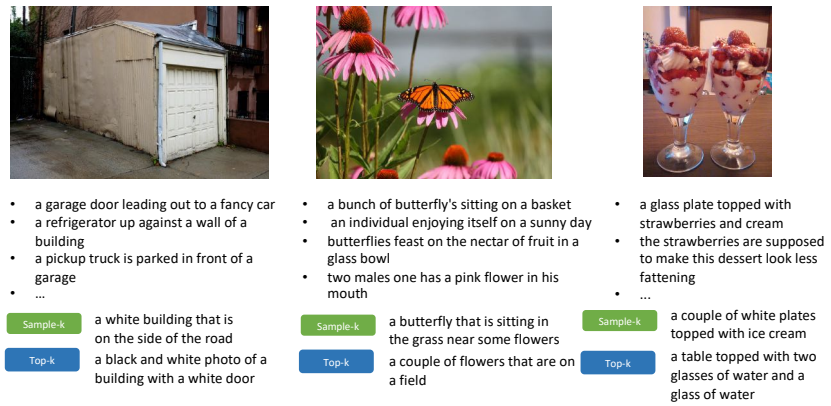


Figure 5.13: Qualitative examples of generated captions on NoCaps **out-domain** samples where the captions retrieved for the given image can be noisy and irrelevant. Here we retrieve four captions for each image.

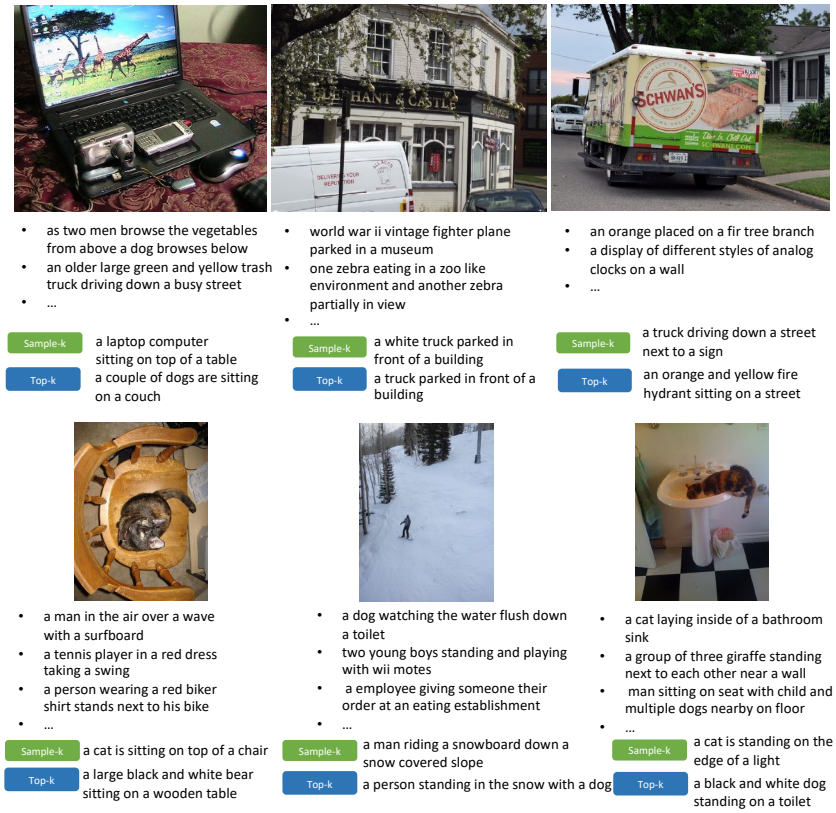


Figure 5.14: More qualitative examples of generated captions when **randomly** retrieving four captions for a given image.

Chapter 6

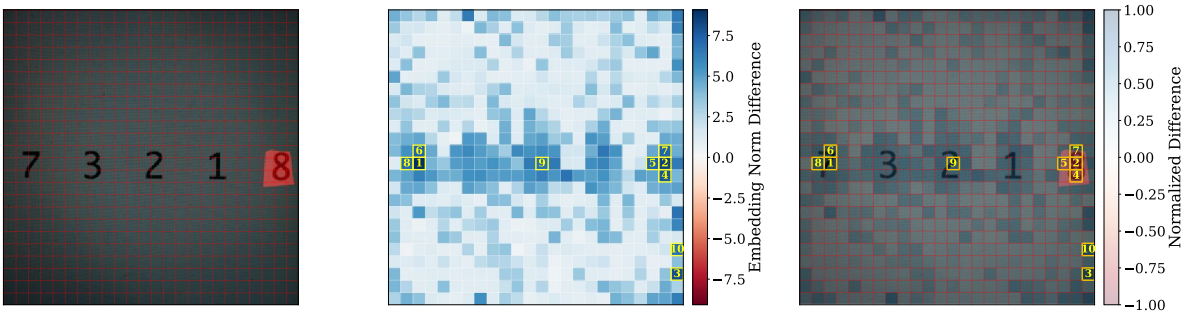
Lost in Embeddings: Information Loss in Vision-Lanugage Models

Abstract

Vision–language models (VLMs) often process visual inputs through a pretrained vision encoder, followed by a projection into the language model's embedding space via a connector component. While crucial for modality fusion, the potential information loss induced by this projection step and its direct impact on model capabilities remain understudied. We propose two novel approaches to quantify such visual information loss in the projection by analyzing the latent representation space. First, we evaluate semantic information preservation by analyzing changes in k-nearest neighbor relationships between image representations, before and after projection. Second, we directly measure information loss by reconstructing visual embeddings from the projected representation, localizing loss at an image patch level. Our experiments reveal that connectors fundamentally alter visual semantic relationships—k-nearest neighbors of the visual embeddings diverge by 40-60% post-projection, correlating highly with degradation in retrieval performance. The patch-level embedding reconstruction provides interpretable insights for model behavior on visual question-answering tasks, finding that areas of high information loss reliably predict instances where models struggle.

6.1 Introduction

Vision–language models (VLMs) have the unique capability of integrating image and language processing. Many of these models employ small modules, known as *connectors*, to bridge the gap between the visual and textual embedding spaces. Tipically, connectors project visual representations into embedding sequences that language models can process



(a) Input image with red answer mask

(b) Embedding norm signed difference

(c) Image overlay with norm difference

Figure 6.1: Visualization of patch-wise information loss in the embeddings explains the incorrect predicted answer in VizWiz Grounding VQA. For the question "What is the fifth number?", LLaVA incorrectly predicted "18". Figure 6.1b display the difference between the L^2 norm of the original and the reconstructed patch embeddings. Blue regions indicate where original embeddings have larger norms than predicted embeddings, while red regions show where predicted embeddings have larger norms. The top 10 high-loss patches are marked by yellow squares. Figure 6.1c shows high loss occurring in several answer-relevant patches contribute to the incorrect prediction.

(Chen et al., 2024a; Liu et al., 2023a; Deitke et al., 2024; Laurençon et al., 2024; Chen et al., 2024b; Zhang et al., 2025; Sun et al., 2024). Common connector architectures include linear layers (Liu et al., 2023a), multi-layer perceptrons (MLPs), or attention-based approaches (Jaegle et al., 2021; Laurençon et al., 2024). While connectors efficiently project rich visual features into embedding spaces compatible with language modules, enabling cross-modal integration (Li and Tang, 2024), this process typically involves dimensionality reduction that may compress important visual information. This raises fundamental questions about the nature and extent of potential *information loss* during projection and whether such loss degrades downstream task performance.

The information loss involved with the connector module may have a detrimental effect on the downstream performance, yet few studies have systematically analyzed it. In fact, most previous work (Lin et al., 2024; Laurençon et al., 2024) has focused on assessing the task-level effect of different connector choices, without actually measuring the loss of visual information linked to the different connector types. Part of the challenge lies in identifying a metric to asses the *visual information* contained in the embeddings before and after the projection.

To address this gap, we present an evaluation framework that quantifies information loss in VLM connector modules from two complementary perspectives: a global *geometric* analysis of how projection alters the morphology of the embedding space, and a localized

assessment of patch-level information preservation. We first measure geometric information loss through careful examination of the structure of latent visual representations. By introducing k**-nearest neighbors overlap ratio**, we can measure how much the neighborhoods of image embeddings change before and after the projection in the latent representation space, thereby estimating how well geometric and semantic relationships are preserved. Second, to measure localized information loss, we train a model to reconstruct the original visual embeddings from the projected embeddings. This **patch-level visual embedding reconstruction** allows us to pinpoint the high-loss regions in the image—areas where visual features are hard to recover after projection (Figure 6.1). This two-step approach provides both quantitative metrics and interpretable visualizations, offering insights into the nature of information transformation during vision-text integration.

6.2 VLMs and Connectors

Integrating visual and textual inputs is fundamental for VLMs to process multimodal information effectively. Existing VLMs typically employ two main approaches (Li and Tang, 2024): models like LLama3.2 (gra, 2024) and BLIP (Li et al., 2023b) leverage cross-modal attention mechanisms, while others such as LLaVA (Liu et al., 2023a) and Qwen-2.5-VL (Bai et al., 2025) adopt connectors to project visual representations into latent vectors compatible with large language models (LLMs).

Lin et al. (2024) categorize connectors into two types: feature-preserving and feature-compressing connectors. Feature-preserving connectors preserve the number of patch embeddings, such as the two-layer MLP connector in LLaVA. In contrast, feature compressing connectors project image patch embeddings to a shorter sequence, including the perceiver sampler in Idefics2 (Laurençon et al., 2024) and the patch merger in Qwen-2.5-VL (Bai et al., 2025). In this paper, we estimate information loss in both types of connectors.

6.2.1 Formalizing Encoders and Connectors

We now define connector-based vision-language models using dependent types. First, we consider the textual input. Let Σ be an alphabet of symbols. A **string encoder**, ϕ, is a function with a dependent type, mapping a string σ to a sequence of real-valued embedding vectors. Formally,

$$\phi\colon \Sigma^N \to (\mathbb{R}^D)^N, \tag{6.1}$$

where $N \in \mathbb{N}$ is a parameter in the dependent type that denotes the length of the input string, and D is the dimensionality of the embedding vectors. This represents a family of functions, one for each N, mapping sequences of N symbols to sequences of N vectors in $\mathbb{R}^D$.

We now turn to the visual input. Let Δ be a set of **image patches**. Each patch $\delta \in \mathbb{R}^{H \cdot W \times C}$ is a 3-dimensional array where H and W represent the height and width dimensions, and C is the number of color channels per pixel. A two-dimensional image of patch dimensions $M_1 \times M_2$ can thus be represented as an element of $\Delta^{M_1 \times M_2}$. Where $\Delta^{M_1 \times M_2}$ denotes the set of all possible $M_1 \times M_2$ grids of patches. The **vision encoder** is formalized as a dependent type:

$$\psi \colon \Delta^{M_1 \times M_2} \to (\mathbb{R}^{D'})^{M_1 \times M_2}, \tag{6.2}$$

where M_1 and M_2 are parameters in the dependent type, representing the grid dimensions of the image patches, and D' is the visual embedding dimension. This maps a grid of image patches to a grid of embedding vectors.

A **connector** module transforms the vision encoder's output to match the dimensionality of the text encoder—projecting visual embeddings of dimension D' to text-compatible dimension D. We define the connector as a function of type:

$$\textsc{conn} \colon (\mathbb{R}^{D'})^{M_1 \times M_2} \to (\mathbb{R}^{D})^{M_C}, \tag{6.3}$$

where we typically have $M_C \leq M_1 M_2$. We also use $\textsc{c}$ as shorthand for $\textsc{conn}$.

For combining the output of the string encoder and the vision encoder, we define a **flattener** that combines visual and textual embeddings into a unified sequence:

$$\textsc{flat} \colon (\mathbb{R}^{D})^{M_C} \times (\mathbb{R}^{D})^{N} \to (\mathbb{R}^{D})^{M_C + N} \tag{6.4}$$

This creates a sequence of length $M_C + N$ by concatenating the flattened grid of visual embeddings with the sequence of text embeddings.

The complete vision–language models we consider can then be expressed as the a composition of these functions:

$$\mathrm{VLM}(x, \sigma) = \mathrm{LM}(\textsc{flat}(\textsc{conn}(\psi(x)), \phi(\sigma))) \tag{6.5}$$

where $x \in \Delta^{M_1 \times M_2}$ is an input image, $\sigma \in \Sigma^N$ is an input text sequence, and LM is an auto-regressive language model that predicts probability of next tokens.

We focus on quantifying the information loss at the connector module defined in Equation 6.3. Formally, the information loss over the connector is a function $\mu : (\psi(x), \textsc{conn}(\psi(x))) \to \mathbb{R}_{\geq 0}$. We explore how such measure correlate and explain model performance.

6.3 Quantifying Information Loss

We propose two methods for quantifying information loss over the projection step described above. The first method quantifies structural preservation of semantic embeddings by

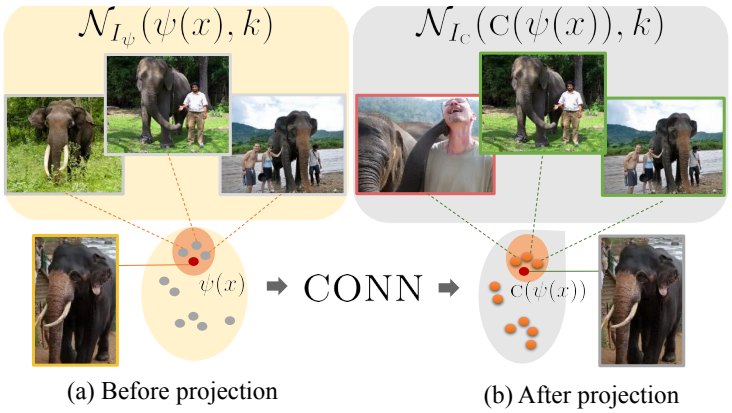


Figure 6.2: The k-nearest neighbors overlap ratio measures the overlap of an image's neighbors before and after projection. In this example, with $k = 3$, the overlap ratio is 0.67 because two out of the three nearest neighbors are identical in both representation spaces.

measuring the overlap between each image representation's k-Nearest Neighbors (k-NN, Fix and Hodges (1951)) before and after projection. Figure 6.2 gives an example where the nearest neighbors overlap but differ in ranking. The second method evaluates patch-level representation (Figure 6.1) distortion by training an *ad hoc* neural network to reconstruct the original image embedding from its projected representation, detailed in Section 6.3.2.

6.3.1 k-Nearest Neighbors Overlap Ratio

To quantify geometric information loss during projection in visual representation spaces, we propose the **k-nearest neighbors overlap ratio (KNOR)**, a measure grounded in the preservation of the k-NN relationship between embedded images before and after projection through the connector. Let I be a finite set of images, ψ a vision encoder, and CONN (C for short) a connector as described in §6.2.1. We use $I_\psi = \{\psi(x)\}_{x \in I}$ to indicate the family of embedded images, and $I_C = \{\text{CONN}(\psi(x))\}_{x \in I}$ for the projection of the embedded images. The k-NN overlap ratio for an image x is defined as

$$\mathcal{R}(x, k) \stackrel{\text{def}}{=} \frac{\left|\mathcal{N}_{I_\psi}(\psi(x), k) \cap \mathcal{N}_{I_C}(C(\psi(x)), k)\right|}{k} \tag{6.6}$$

Where $\mathcal{N}_{I_\psi}(\psi(x), k)$ is the set of k-nearest neighbors of $\psi(x)$ among the pre-projected embeddings, and $\mathcal{N}_{I_C}(C(\psi(x)), k)$ is the set of k-nearest neighbors of $C(\psi(x))$ among the projected embeddings. The **average overlap ratio** is given by

$$\overline{\mathcal{R}}(k) \stackrel{\text{def}}{=} \frac{1}{|I|} \sum_{x \in I} \mathcal{R}(x, k) \tag{6.7}$$

The average overlap ratio measures how well the local geometric structure is preserved after projection. Lower overlap ratio corresponds to more geometric information loss due to projection, while higher overlap suggests faithful retention.

6.3.2 Embedding Reconstruction

While KNOR quantifies the loss of geometric relationships between image embeddings, it cannot detect loss of patch-level visual features. To address this, we further quantify and localize patch-level information loss by attempting to reconstruct the original vision embeddings from their projected representations.

Specifically, given a connector CONN defined in Equation 6.3 and set of images $I \subset \Delta^{M_1 \times M_2}$, we train a **reconstruction model** $f_\theta : (\mathbb{R}^D)^{M_C} \to (\mathbb{R}^{D'})^{M_1 \times M_2}$ to minimize reconstruction loss. For each patch index $(i, j) \in M_1 \times M_2$, we define the per-patch loss as

$$\mathcal{L}_{\text{patch}}(x, i, j) \stackrel{\text{def}}{=} \|\psi(x)_{(i,j)} - f_\theta(\text{C}(\psi(x)))_{(i,j)}\|_2^2 \tag{6.8}$$

which measures the squared Euclidean distance between the original vision embedding and its reconstruction for each patch. The total reconstruction loss is therefore the sum of the patch-wise losses across all patches and images:

$$\mathcal{L}_{\text{recon}}(I) \stackrel{\text{def}}{=} \sum_{x \in I} \sum_{\substack{(i,j) \in \\ M_1 \times M_2}} \mathcal{L}_{\text{patch}}(x, i, j) \tag{6.9}$$

This patch-wise reconstruction enables us to identify and visualize the spatial distribution of information loss across the image.

6.4 Experimental Setup

We quantify information loss using both methods across three open-weights connector-based vision-language models on six datasets spanning question answering, captioning, and retrieval tasks. We assume that greater structural and semantic information loss during projection through the connector leads to reduced neighborhood overlap, while greater patch-wise information loss results in higher reconstruction error.

6.4.1 Pretrained VLMs

We consider three VLMs including LLaVA (Liu et al., 2023a), Idefics2 (Laurençon et al., 2024), and Qwen-2.5-VL (Bai et al., 2025). LLaVA uses a two-layer MLP as the connector, preserving total number of patches for each image. In contrast, Idefics2 uses an attention-based perceiver resampler (Jaegle et al., 2021) that projects image embeddings to a

fixed-length sequence of embeddings. Qwen-2.5-VL uses a MLP-based patch merger which merges every four neighboring patch representations into one. We use the 7B-instruct model variants for LLaVA and Qwen-2.5-VL, and the Idefics2-8B-instruct model.

6.4.2 Evaluation Datasets

We evaluate on six diverse datasets, each of which probes different aspects of visual understanding.

SEED-Bench (Li et al., 2024b) provides categorized multiple-choice questions spanning cognitive tasks from basic scene understanding to complex visual reasoning.

VizWiz Grounding VQA (Chen et al., 2022) includes real-world visual assistance scenarios with grounding-based question answering.

VQAv2 (Goyal et al., 2017) covers open-ended questions that test general visual comprehension.

CUB-200-2011 (Wah et al., 2011) is a commonly used dataset for fine-grained image retrieval that covers 200 species of birds.

Flickr30k (Young et al., 2014) and **COCO** Karpathy test set (Karpathy and Fei-Fei, 2017) are used for image captioning evaluation.

Together, these datasets offer complementary perspectives on how different types of visual information are preserved during projection and how information loss impacts various downstream tasks.

6.4.3 Embedding Reconstruction Models

We build models to reconstruct image patch embeddings from connector outputs. These reconstruction models are intentionally designed with larger capacity than the original connectors, including expanded hidden dimensions and additional hidden layers. This controlled setup ensures our models are trained to recover the original visual representations without creating new bottlenecks in the reconstruction process.

Architecture We tailor our reconstruction models to each VLM's connector architecture. For LLaVA, which preserves the number of image patches during projection, we use a simple three-layer MLP with a 2048-dimension hidden layer. For Idefics2 and Qwen-2.5-VL, which compress sequence length from $M_1 \times M_2$ to M_C, we implement transformer-based models to handle the differences in sequence length. The reconstruction model projects connector outputs to hidden embeddings with positional encodings before processing them through a 16-layer, 16-head transformer encoder with 2048-dimensional vectors. Table 6.1 summarizes the parameters of the reconstruction models and their input and output dimensions. Please see Appendix 6.9.3 for ablation analysis on the reconstruction model structure.

Model	$M_1M_2 \times D'$	$M_C \times D$	**|conn|**	$\|f_\theta\|$
LLaVA	576×1024	576×4096	21M	27M
Idefics2	576×1152	64×4096	743M	844M
Qwen-2.5-VL	576×1280	144×3584	45M	843M

Table 6.1: Model parameters and embedding dimensions. |CONN| denotes number of parameters in the connector and $|f_\theta|$ represents number of parameters of the reconstruction model. Pre- and post-projection embedding dimensions are listed as $M_1M_2 \times D'$ and $M_C \times D$.

Training We train each of the embedding reconstruction models on the COCO 2017 train set (Lin et al., 2014) for 30 epochs with early stopping. We apply a learning rate of $1e{-}4$, dropout of 0.1, and a total batch size of 128. For training stability, we apply normalization to both pre- and post-projection embeddings using mean and standard deviation of the dataset.

6.5 Neighbor Rankings and Semantic Information are Not Preserved

We calculate KNOR (Section 6.3.1) for images in the SeedBench validation set, a subset of the VQAv2 validation set with $10,000$ images, and the validation set of Vizwiz grounding VQA dataset. It is intuitive that higher neighborhood overlap ratios suggest that the projection better preserves the relationships between image embeddings. As the neighborhood rankings directly impact image retrieval tasks, we also evaluate retrieval performance on the CUB dataset using both pre- and post-connector visual embeddings.

6.5.1 Low Overlap Ratio for All Models

In Figure 6.3, we show the neighborhood overlap ratio across $k = 10$, 50, and 100 nearest neighbors, averaging through all unique images in the evaluation datasets.[1] We can observe that the neighborhood overlap ratios are around 50% for LLaVA and Idefics-2, with LLaVA achieving 61.6% overlap as the maximum when considering 100 nearest neighbors, whereas Qwen-2.5-VL lost almost 90% of the neighborhood ranking information. This suggests a significant reordering of nearest neighbors post-projection across all models. However, even LLaVA shows notable neighbor reshuffling, especially at smaller neighborhood sizes (k=10).

[1]Visual embeddings pre- and post-connector projection have a 1-1 mapping to the input image, and these visual embeddings are not impacted by the language model prompts.

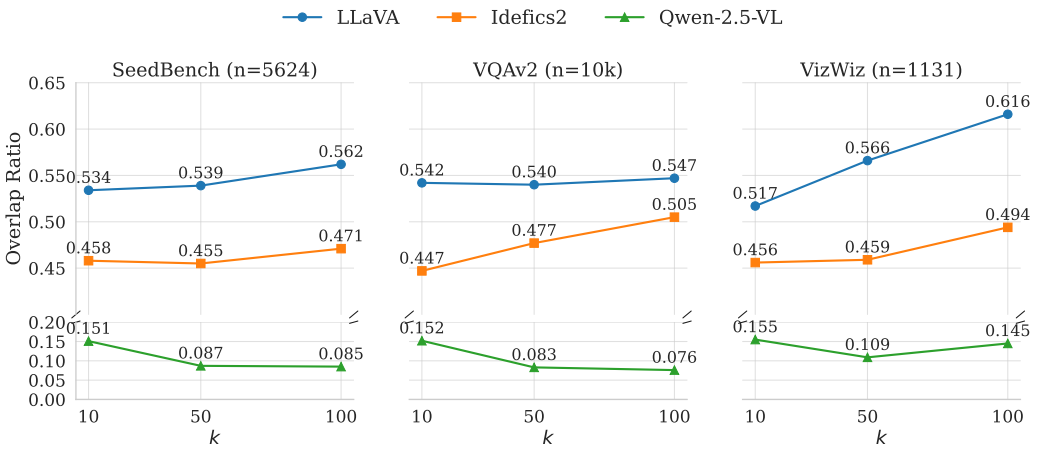
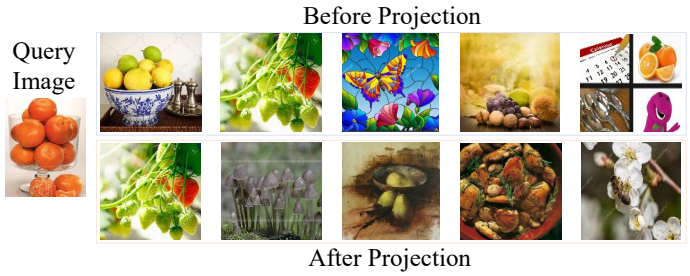


Figure 6.3: Neighborhood overlap ratios across three datasets: SeedBench validation, a 10,000-sample subset of VQAv2 validation, and Vizwiz grounding VQA validation. Analysis using 10, 50, and 100 nearest neighbors shows overlap ratios below 0.62 for all models, suggesting connectors poorly preserve geometric relationships and neighbor rankings for the visual representations.

In Figure 6.4, we visualize the nearest neighbors of a given query image, revealing significant neighbor reordering across all models. However, for Qwen-2.5-VL, the neighbors obtained with post-projection embeddings are more semantically similar to the query image. We suspect that this phenomenon could stem from its continuous training of the image encoder in the pretraining stage and the patch merging, which yields more semantically meaningful post-projection embeddings. Other VLMs such as LLaVA use a frozen vision encoder, where the connector is updated to inherit features from the pretrained encoder. However, in Qwen-2.5-VL, continued pretraining with an unfrozen vision encoder produces fundamentally different learned visual embeddings. This indicates that the pre- and post-projection visual representations are not equivalent, but may not necessarily lead to worse semantic representations of the image.

6.5.2 Image Retrieval Evaluation

To verify if structural information loss correlates with a degradation in the semantic representation of images, we evaluate on the CUB-200-2011 image retrieval test set (Wah et al., 2011). We perform zero-shot image retrieval with pre- and post-connector embeddings for each query image, excluding the query image itself from the gallery. The pre-and post-projection embeddings are indexed with FAISS (Douze et al., 2024), and we experiment with retrieving similar images based on both the L^2 distance and the inner product similarity (Table 6.8 in Appendix) of the image representations.



(a) Five nearest neighbors of LLaVA image embeddings

(b) Five nearest neighbors of Idefics2 image embeddings

(c) Five nearest neighbors of Qwen-2.5-VL image embeddings

Figure 6.4: Comparison of five nearest neighbors searched with pre-projection (top) and post-projection (bottom) embeddings using different models. The first image in each row is the query image, followed by its nearest neighbors. For Qwen-2.5-VL, despite a low neighborhood overlap ratio, post-projection embeddings retrieve more semantically similar images.

We report the recall scores at rank 1 (R@1) and rank 5 (R@5) in Table 6.2. Consistent with our observations from the neighborhood overlap visualization (Figure 6.4), we observe semantic degradation of 41.4% and 18.8% of R@5 for LLaVA and Idefics model, respectively. In contrast, for the Qwen-2.5-VL model, the improved image retrieval performance with

Model	Emb	Recall		Correlation	
		R@1	R@5	R@1	R@5
LLaVA	Pre	8.34	21.82	0.05	0.08
	Post	6.16	17.22	0.11	0.11
Idefics2	Pre	**13.10**	**30.81**	0.19	0.23
	Post	10.87	25.28	**0.22**	**0.28**
Qwen-2.5-VL	Pre	4.23	11.74	0.10	0.13
	Post	10.65	26.44	0.16	0.21

Table 6.2: Zero-shot retrieval performance on CUB test set using L^2 for similarity measure. R@k denotes Recall at rank k. We calculate the Spearman correlation scores with R@k and the average overlap ratio considering 100 nearest neighbors. p values are smaller than $1e{-}5$ for all correlation scores.

post-projection embeddings suggests that the low overlap ratio stems from the substantial differences between the two sets of visual embeddings, with the post-projection embeddings capturing more semantic features. We also observe positive correlation between the k-NN overlap ratio and the retrieval R@1 and R@5 scores for all models. The correlation is more significant especially when using post-projection embeddings. This suggests that our proposed k-NN measure correlates with performance on tasks requiring fine-grained visual discrimination.

6.6 Reconstruction and Model Behavior

Beyond KNOR reflecting semantic and geometric losses, we examine patch-level information loss by reconstructing visual representations $\psi(x)$ from their projections $\text{CONN}(\psi(x))$ (Equation 6.8). Higher reconstruction loss indicates greater information loss. This patch-level loss measure enables precise localization of visual feature degradation.

6.6.1 Reconstruction Loss Impacts Captioning

Our embedding reconstruction evaluation follows two steps: 1) we train a reconstruction model for each VLM using paired pre- and post-projection embeddings from images in the COCO 2017 train set (as described in Section 6.4.3); 2) we apply these reconstruction models to predict the original image representations from their projected counterparts.

[2]We notice Qwen-2.5-VL is particularly sensitive to the task prompt; here we use the prompt suggested in the original paper (Bai et al., 2025).

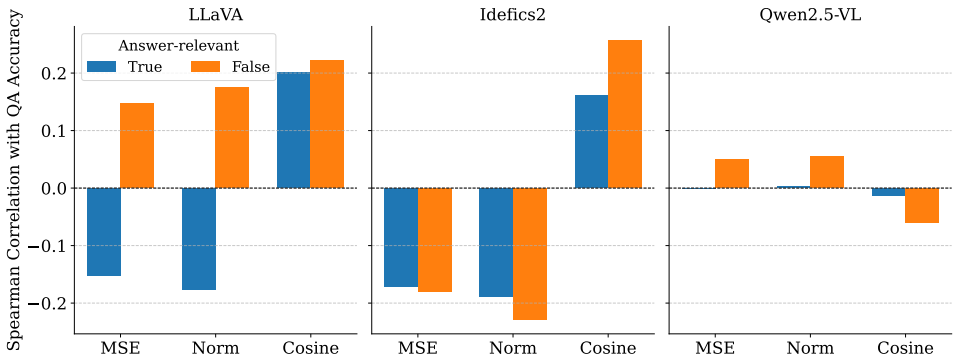


Figure 6.5: Correlation between reconstruction loss and question-answering accuracy on the VizWiz grounding VQA task. For LLaVA and Idefics2, all correlations have a p-value $< 5e{-}5$, indicating statistically significant relationships, whereas no clear correlation is observed for Qwen-2.5-VL. The reconstruction loss occurs in both answer-relevant and irrelevant patches. Loss in relevant patches negatively affects performance of LLaVA and Idefics2. "Norm" represents differences between the L^2 norm of the embeddings.

For image captioning, we measure the reconstruction loss for images in the Flickr30k validation set and COCO Karpathy test split. We use CIDEr score (Vedantam et al., 2015) to evaluate the quality of the generated captions. Table 6.3 summarizes the overall average reconstruction loss of the three models on the captioning test datasets. For both datasets, we observe lower average reconstruction loss yields better captioning performance. We also investigate how reconstruction loss impacts captioning for each individual image by calculating the correlation between per-sample CIDEr score and reconstruction loss per-image. In Table 6.4, the spearman correlation indicates higher reconstruction loss for a given image corresponds to worse captioning for Idefics and LLaVA, indicating by the negative correlation with p values smaller than $1e{-}5$. Please see more visualization in Figure 6.11. For Qwen-VL, we did not observe obvious correlation for individual images. The large gap of CIDEr scores between the highest and lowest reconstruction loss samples for LLaVA and Idefics2 suggests substantial impact on downstream tasks.

6.6.2 Loss at Patch-level Visual Features Explains Question Answering Behaviors

To further distinguish whether the reconstruction loss stems from selective feature preservation or actual information loss, we visualize the patch-level loss for images in the VizWiz grounding VQA validation dataset. This dataset is particularly suitable for our analysis as it provides answer grounding—binary masks indicating image regions relevant to each question. By examining the relationship between the reconstruction loss for the

Model	COCO	Flickr30k
Reconstruction loss (avg / std)		
LLaVA	0.087 / 0.016	0.097 / 0.019
Idefics2	0.796 / 0.082	0.854 / 0.074
Qwen-2.5-VL	1.069 / 0.117	1.069 / 0.115
Overall CIDEr Scores		
LLaVA	81.28	56.79
Idefics2	53.64	39.22
Qwen-2.5-VL	13.04	12.85

Table 6.3: Reconstruction loss on COCO and Flickr30k test sets. Top: reconstruction loss averaged over all samples, where LLaVA achieves lowest reconstruction error. Bottom: CIDEr scores of zero-shot captioning.[2]For both datasets, we observe better overall captioning performance with lower average reconstruction loss.

answer-relevant image patches and question-answering accuracy, we can assess whether the projection preserves task-relevant visual information.

We report the Spearman correlation between the reconstruction loss and the question answering accuracy in Figure 6.5. For LLaVA, we observe a negative correlation between prediction accuracy and reconstruction loss in answer-relevant patches, while a positive correlation is found in irrelevant patches. This indicates that information loss in answer-relevant patches negatively impacts model performance, whereas loss in irrelevant patches has a less significant effect. For Idefics2, we can see that information loss in any patches would hurt question answering accuracy. We do not observe significant correlation for Qwen-2.5-VL, which is consistent with our findings in the captioning tasks.

As shown in Figure 6.1, identifying distorted features allows us to pinpoint visual information that becomes inaccessible or less reliable for the language model. For instance, reconstruction loss in the patches of the fifth number "8" rank among the top ten of all image patches, suggesting that the model may have struggled to answer the question due to lost details necessary for identifying the number. This analysis introduces a new visualization approach to examine VLM limitations, particularly in scenarios requiring reasoning or recognizing fine-grained viusal features. Please see more visualization examples in Appendix 6.9.5.

6.7 Related Work

A series of analyses has been conducted to investigate the modality gap and representation limitations of contrastive-based VLMs (Schrodi et al., 2024; Liang et al., 2022; Tong et al.,

Model	COCO	Flickr30k
CIDEr Scores for High Loss / Low Loss samples		
LLaVA	73.98 / 86.96	51.79 / 61.74
Idefics2	40.84 / 66.13	29.24 / 53.22
Qwen-2.5-VL	12.45 / 13.56	13.15 / 12.35
Spearman Correlation (ρ / p)		
LLaVA	−0.077 / 0.000	−0.096 / 0.000
Idefics2	−0.214 / 0.000	−0.226 / 0.000
Qwen-2.5-VL	0.001 / 0.975	0.027 / 0.403

Table 6.4: Top: The comparison of CIDEr scores for top 25% highest and 25% lowest reconstruction loss samples, reported as "High Loss / Low Loss" Bottom: Spearman correlations (ρ) of per-sample reconstruction loss and captioning CIDEr scores.

2024). These studies reveal that the representational shortcomings in CLIP embeddings subsequently impact the visual perception capabilities of VLMs relying on such vision encoders. For connector-based VLMs, Zhang et al. (2024) demonstrates that the latent space sufficiently retains the information necessary for classification through probing across different layers, and Lin et al. (2024) demonstrates the impact of different connectors on VLMs' downstream performance. However, there remains a significant gap in understanding whether fine-grained visual information, crucial for tasks such as visual grounding (Krishna et al.) and question answering (Chen et al., 2022), is lost in the process. In this paper, we focus on the connector-based models to understand the information transformation. To the best of our knowledge, our paper is the first to directly quantify information loss of the connectors from the representation perspective, offering deeper insights into where and what specific information is lost from the visual features.

6.8 Conclusion and Future Work

Our study systematically evaluates information loss during visual-to-language projection in VLM connectors through two key metrics: neighborhood overlap ratios and embedding reconstruction. Our quantitative framework captures two critical aspects of the information loss 1) significant structural shifts in global semantic relationships shown by 40-60% divergence in nearest-neighbor rankings, and 2) patch-level reconstruction loss that correlates with degraded performance in captioning and fine-grained visual QA tasks. Our patch-level reconstruction also enables visualization of local information loss, offering interpretable explanations for model behaviors.

Our findings suggest two key properties of an effective connector: 1) preserving or improving semantic representation of images, and 2) preserving visual information most relevant to the text context. These findings could guide further improvements in VLM connectors. For example, the reconstruction loss at the embedding level could potentially be incorporated during model pretraining as regularization. Future work could also explore designing dynamic projection layers or better visual feature selection mechanisms for modality fusion.

Ethics Statement

We foresee no ethical concerns with our research project. In particular, ours is merely a scientific study of VLMs and provides no artifacts that can be used in a real-world scenario.

Limitations

In this study, we evaluate the information loss introduced by connectors in VLMs. However, several limitations should be noted. First, due to variations in model architectures and pretraining strategies, our findings may be specific to the connector-based VLMs analyzed and may not generalize to architectures that employ cross-attention for modality fusion. Second, our experiments focus on connectors in VLMs within the 7B–8B parameter range. Expanding the analysis to models of different sizes could provide deeper insights into the relationship between model scale and information loss. Third, our pixel-level reconstruction experiments (Appendix 6.9.6) yielded inconclusive results in quantifying information loss, possibly due to limitations in our chosen image generation model and training dataset size. Additionally, while we empirically validate our k-NN overlap ratio and embedding reconstruction metrics, a formal theoretical characterization would further strengthen their reliability. Finally, our reconstruction experiments cannot conclusively determine whether the observed information loss stems from the connector layer itself or from potential learning limitations of the trained reconstruction network.

6.9 Appendix

6.9.1 Connectors in Autoregressive Vision-Language Models

Idefics2 Idefics2 leverages a perceiver resampler (Jaegle et al., 2021) as the connector. The perceiver resampler forms an attention bottleneck that encourages the latent representations to attend to the most relevant inputs in a high-dimensional input array through

iterative cross-attention layers. In other words, the cross-attention module projects the high-dimensional inputs into a fixed-dimensional learned representation. Please refer to Laurençon et al. (2024) for more details.

LLaVA LLaVA (Liu et al., 2023a) uses a two layer MLP to project the image embeddings to the language model's embeddings space. The MLP projector preserves the image feature length – number of patches extracted by the image encoder.

Qwen-2.5-VL Qwen-2.5-VL (Bai et al., 2025) uses a patch merger (two-layer MLP) to reduces the length of the input image features. The image representations of the neighboring four patches in the image are first merged, and then passed through a two-layer MLP to project the image representation to the LM embedding dimension.

6.9.2 Procrustes analysis

We also attempt to find the optimal geometrical transformation from the post-projection embedding space to the pre-projection one through Procrustes analysis (Gower, 1975) – a method often used for supervised alignment of embeddings (Artetxe et al., 2018). The alignment error reflects the degree of structural similarity of the two embedding spaces.

We use mean-pooled image embeddings from LLaVA, Idefics2, and Qwen-2.5-VL. As the pre- and post-projection embeddings have different embedding dimensions and sequence lengths, our analysis follows three steps to complete the embedding alignment. We first take the mean-pooled image representation by averaging over the sequence length, producing fixed-size vectors of size D' and D. We then use PCA (Hotelling, 1933) on the mean-pooled post-projection embeddings to project them to the same dimension of the mean-pooled pre-projection embeddings.

Orthogonal transformation matrix $\mathbf{R}$ was derived through singular value decomposition of the cross-covariance matrix $\bar{X}^\top \bar{T}$, where $\bar{X} \in \mathbb{R}^{D'}$ represents mean-pooled pre-projection embeddings and $\bar{T} \in \mathbb{R}^{D'}$ the PCA-transformed post-projection embeddings. Then the orthogonal transformation matrix is learned to best align these two sets of embeddings by minimizing the Euclidean distance. The reconstruction error are reported in Table 6.5. Figure 6.6 visualizes the alignment of LLaVA embeddings through procrustes analysis.

Our analysis reveals fundamental limitations in linear alignment of the image embeddings. The high alignment errors of 16.62 for LLaVA and 4.41 for Qwen-2.5-VL indicate the inherent difficulty of preserving geometric relationships through rigid transformations. While serving as a critical baseline for structural fidelity assessment, this constrained linear approach explains why our proposed non-linear embedding reconstruction approach achieves significantly lower errors.

In Figure 6.6, we visualize the alignment for LLaVA pre- and post-projection embeddings, as well as the embeddings learned through the linear transformation learned. From

Model	Mean	Std	Min	Max
LLaVA	16.62	3.16	8.76	23.65
Idefics2	4.93	0.08	4.78	5.70
Qwen-2.5-VL	4.41	0.09	4.24	5.05

Table 6.5: Procrustes analysis results. We report the alignment error on SeedBench image representations before and after connector projection.

the visualization we can observe that the linear transformation is not able to align the pre- and post-projection embeddings well.

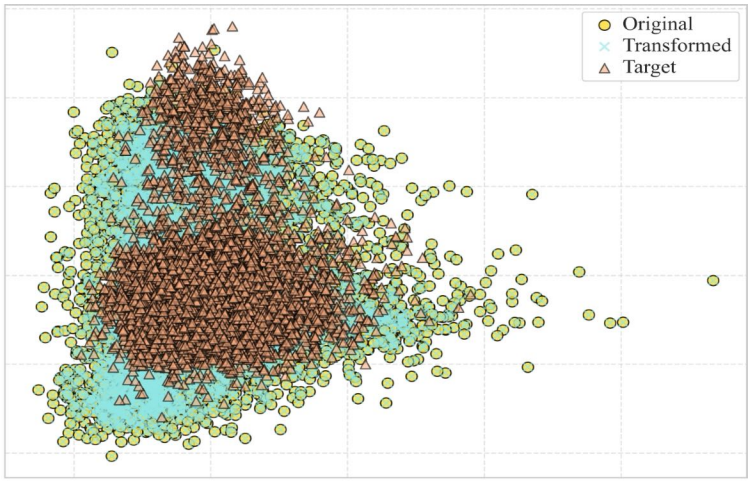


Figure 6.6: Alignment visualization for LLaVA pre- and post-projection embeddings through PCA.

6.9.3 Ablation Studies

Ablation on Reconstruction Model Size and Structure We train three reconstruction models of different sizes for LLaVA: a 27M three-layer MLP, a 39M five-layer MLP, and a 40M Transformer. In Table 6.6, we observe that the 27M model is sufficient for reconstructing LLaVA visual embeddings, and a larger model does not yield better validation loss.

Model	Size	VizWiz	SeedBench	FoodieQA
MLP	27M	Avg 0.050 Std 0.013	0.056 0.011	0.051 0.007
MLP	39M	Avg 0.064 Std 0.015	0.070 0.013	0.065 0.0075
Transformer	40M	Avg 0.237 Std 0.019	0.231 0.025	0.228 0.014

Table 6.6: Reconstruction loss with different architectures across VizWiz, SeedBench, and FoodieQA datasets. Reported values include average loss (Avg) and standard deviation (Std).

Ablation on Index Method for k-NN Overlap Ratio We evaluated k-NN overlap ratio using three different embedding types as search indices: original embeddings, mean-pooled image embeddings, and normalized embeddings (Table 6.7). Since the performance differences were minimal, we selected mean-pooled embeddings for both pre- and post-projection image representations in calculating k-NN overlap ratios.

Overlap Ratio	Index Type					
	IndexFlatL2		IndexFlatL2 (mean pooling)		IndexFlatIP (normalized vectors)	
	mean	**std**	**mean**	**std**	**mean**	**std**
top100	0.466	0.122	0.563	0.107	0.504	0.129
top50	0.488	0.128	0.556	0.120	0.425	0.142
top10	0.490	0.149	0.551	0.160	0.377	0.161
Vector Size						
Before projection	576×1024		1×1024		576×1024	
After projection	576×4096		1×4096		576×4096	

Table 6.7: Ablation on KNN results when using original embeddings, mean pooled image embeddings, and normalized embeddings. We chose to use the mean-pooled embeddings for efficiency due to large embeddings size.

6.9.4 Additional Evaluation Results

CUB image retrieval performance In Table 6.8, we show the complete image retrieval performance on CUB test set using L^2 and inner product for similarity measure. The performance are consistent regardless of the index method used.

Reconstruction loss on VQA datasets For visual question answering tasks, we measure the reconstruction loss for images in the validation set of VizWiz grounding VQA, Seed-Bench, and FoodieQA. Table 6.9 presents overall reconstruction loss. Among all tested models, LLaVA's projected embeddings maintain the highest reconstruction fidelity. The overall reconstruction loss reflects the overall difficulty of recovering information encoded in the visual representations.

Model	**L2**		**IP**	
	R@1	**R@5**	**R@1**	**R@5**
Pre-projection				
LLaVA	8.34	21.82	9.46	24.78
Idefics2	**13.10**	**30.81**	**13.38**	**30.98**
Qwen-2.5-VL	4.23	11.74	6.83	24.23
Post-projection				
LLaVA	6.16 ↓	17.22 ↓	5.54 ↓	20.49 ↓
Idefics2	**10.87** ↓	25.28 ↓	**10.99** ↓	25.15 ↓
Qwen-2.5-VL	10.65 ↑	**26.44** ↑	8.26 ↑	**26.70** ↑

Table 6.8: Zero-shot retrieval performance on CUB test set using L^2 distance and inner product for similarity measure. R@k denotes Recall at rank k. Arrows indicate performance change direction after projection.

6.9.5 Visualization

Patch-level Loss Visualization for Vizwiz Grounding VQA In Figure 6.7, we visualize additional examples of high reconstruction loss patches that contributes to model's failure on answering questions that requires recognizing text in the objects.

Visualization of Neighborhood Reordering In Figure 6.10, we present more k-NN examples on comparison of searching with pre-projection (top) v.s. post-projection (bottom) embeddings.

Dataset	MSE	LLaVA	Idefics2	Qwen-2.5-VL
VizWiz	Avg	**0.115**	0.907	1.069
	Std	0.086	0.298	0.684
SeedBench	Avg	**0.106**	0.872	1.069
	Std	0.071	0.307	0.610
FoodieQA	Avg	**0.113**	0.918	1.069
	Std	0.057	0.283	0.673

Table 6.9: Embedding reconstruction loss of images in the VizWiz, SeedBench, and FoodieQA datasets. We report both average loss (avg) and standard deviation (std). LLaVA's visual embeddings exhibit lowest reconstruction error among all models. The reconstruction performance is consistent to what we have observed for the images in COCO and Flickr30k.

Visualization of reconstruction loss and captioning performance In Figure 6.11 we show visualization of captioning where details in the high-loss patches are missed or inaccurate in the generated caption.

6.9.6 Image Reconstruction with Different Embeddings

Beyond neighbor-overlapping and embedding reconstruction, we aim to investigate how information loss manifests in the reconstructed images themselves. To explore this, we project different representations of visual features onto the input embedding space of a powerful image decoder to assess their reconstruction quality. However, image reconstruction performance depends on various factors, including the expressiveness of the image decoder. As such, this section serves as a preliminary exploration, and we encourage future work in this direction.

For our experiments, we use a fine-tuned VAE decoder[3], trained on the original VAE checkpoint from Stable Diffusion, trying to alleviate the influence of the decoder as a limiting factor in reconstruction quality. To align the sequence length between the vision encoder in the VLM and the expected input length of the VAE decoder, we employ a 6-layer Transformer encoder-decoder module with 4 attention heads. We train the aligner module on the COCO 2017 training set for 100 epochs with three objectives: 1) Embedding loss minimizing the difference between the VAE encoder embeddings and the aligned embeddings from the VLM's visual encoder; 2) Reconstruction loss measuring the mean squared error (MSE) between the original and reconstructed images; 3) Latent loss

[3] https://huggingface.co/stabilityai/sd-vae-ft-mse

quantifying the divergence between the mean and variance of the Gaussian distribution for diffusion.

For the VLM, we use the LLaVA model in our experiments. We evaluate reconstruction performance on both an in-distribution image from the COCO 2017 dev split and an out-of-distribution image, as shown in Figure 6.12. When using embeddings before projection, the overall pixel-wise MSE reconstruction loss is 0.2128, compared to 0.2443 after projection. Figure 6.12 illustrates the reconstructed images for both cases, where pre-projection embeddings yield similar contour preservation with post-projection embeddings.

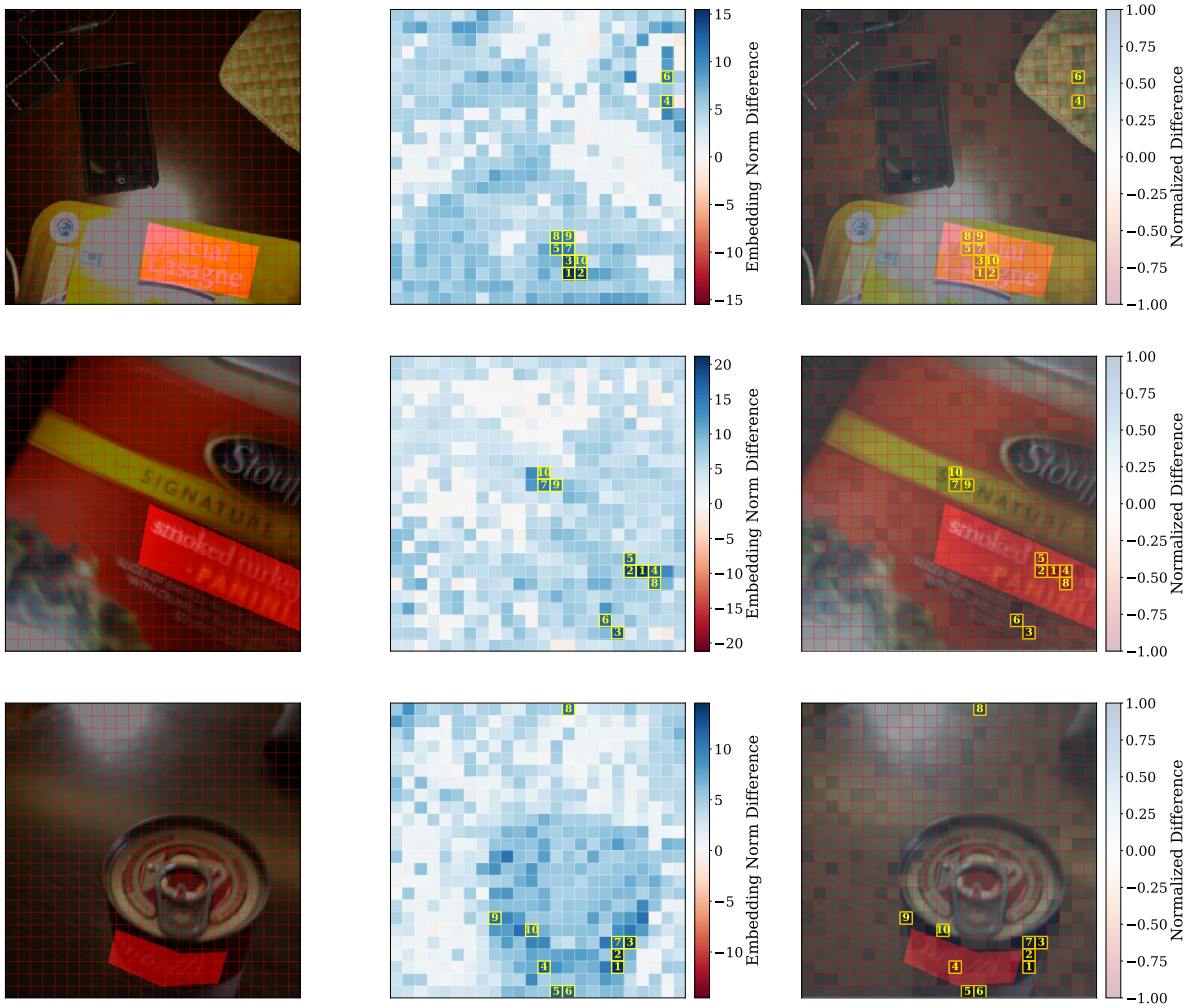


Figure 6.7: Additional visualization of high reconstruction loss patches that contributes to model's failure on answering questions that requires recognizing text in the objects. Left: input images with answer-relevant regions in red masks. Middle: signed difference between post-projection embeddings norms and pre-projection embedding norms. Right: normalized norm differences overlay with the input image, with highest loss patches marked in yellow.

Figure 6.8: Idefics high kNN overlap ratio example, where we can observe the reordering among semantically similar vision embeddings.

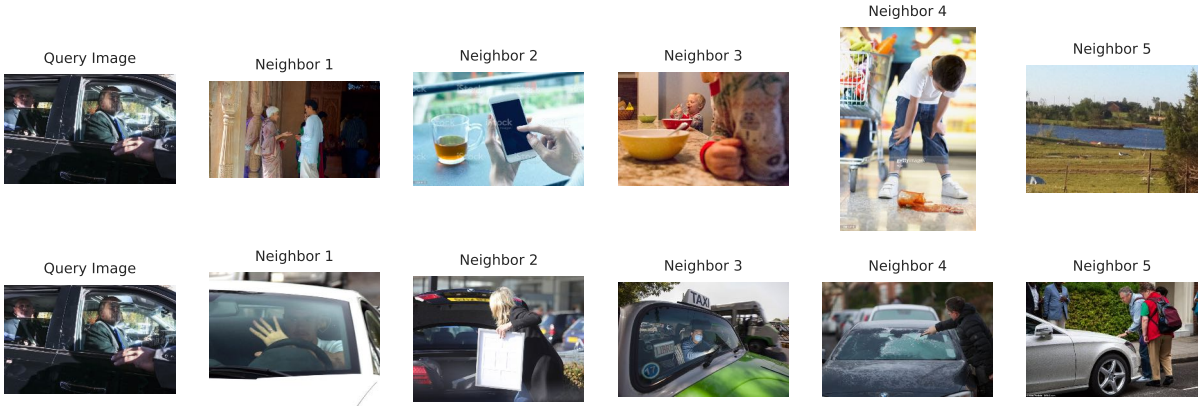


Figure 6.9: Qwen kNN example where the post-projection embeddings are better at retrieving semantically similar images (bottom).

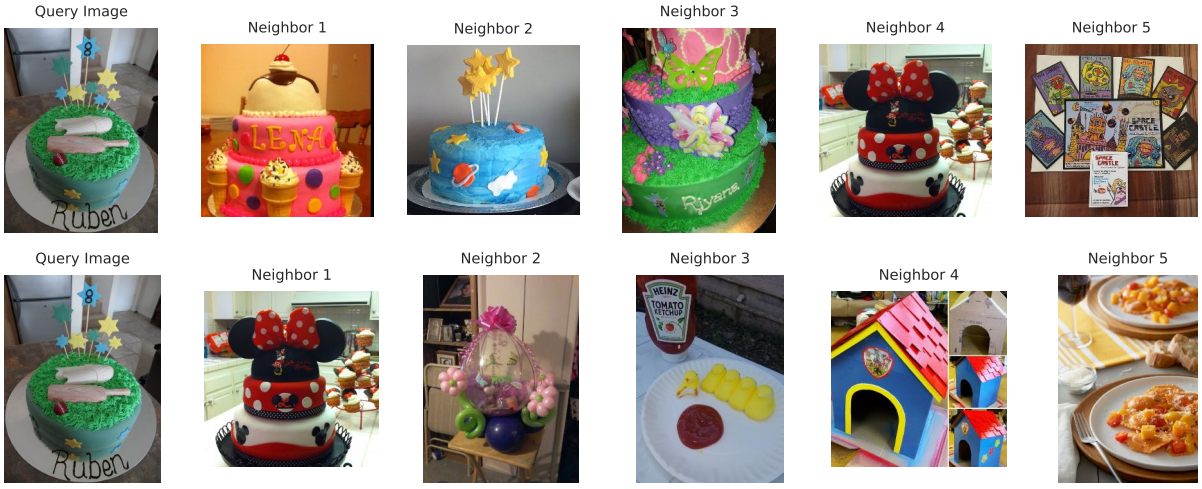
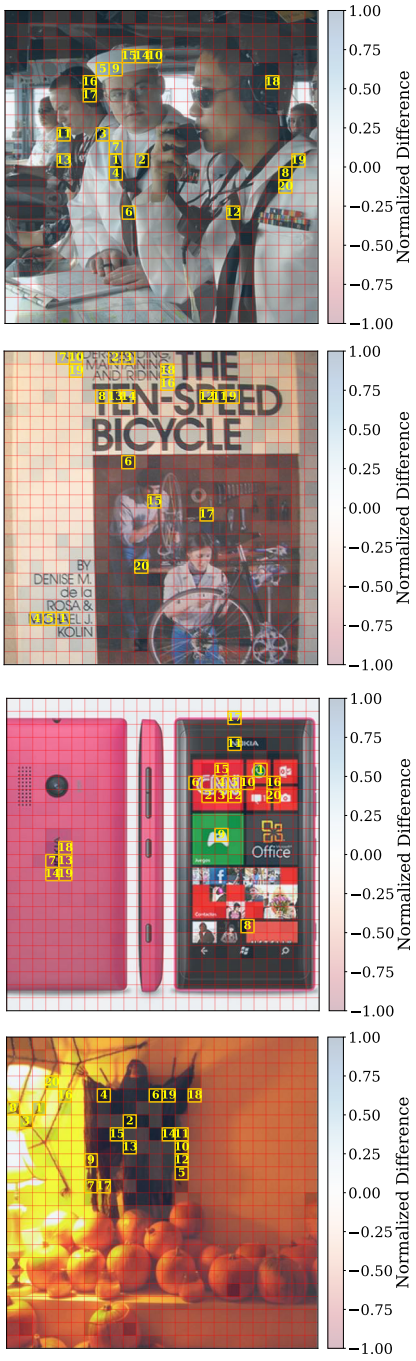


Figure 6.10: LLaVA low kNN overlap ratio example. We can observe the degradation in post-projection embedding.



Reference Captions
- People in **navy** uniforms and one person talking on a walkie- talkie.
- Group of **sailors** in command center with one talking on walkie talkie.

Generated Caption
- A group of men in uniforms are sitting at a table.

Reference Captions
- A book about understanding and maintaining a **ten-speed bicycle**.
- A sign explaining the components of **a 10 speed bike.**

Generated Caption
- A man and a woman are working on a bicycle.

Reference Captions
- Various angle shots of the **Nokia** Windows cell phone.
- A pink smartphone with Windows 8 on the screen.

Generated Caption
- A pink cell phone is displayed next to a red screen.

Reference Captions
- The Halloween display includes a **spiderweb** and lots of pumpkins.
- Multiple pumpkins and a skeleton on the wall.

Generated Caption
- A spooky Halloween display features a witch figure and a bunch of pumpkins.

Figure 6.11: Visualization of low CIDEr score captioning samples and the reconstruction loss overlay with the input image. We can observe that details regarding the high loss patches are missing from the generated captions. High loss patches are marked in yellow squares.

(a) Original (b) Reconstruction with Pre-projection Embeddings (c) Reconstruction with Post-projection Embeddings

Figure 6.12: Image reconstruction with LLaVA pre-and post-projection embeddings on out-of-distribution (top) and in-distribution (bottom) examples.

Chapter 7

Conclusions

The goal of this dissertation is to provide systematic understanding of the pipeline of multimodal models. We started by examining the data variations in training datasets, understanding model learning dynamics and improve learning data efficiency by utilizing different curation methods, including applying existing models such as text-to-image generation models in a model-in-the-loop manner. Recognizing the diverse scenarios of multimodal model applications, we then look at the data variation and diversity at inference time, evaluating state-of-the-art models within cultural contexts by establishing a fair and challenging fine-grained VQA benchmark. Finally, the dissertation examines the vision-language models themselves, interpreting how they represent and fuse multimodal information to make decisions and identifying key bottlenecks in modality fusion components and retrieval-augmented structures.

Together, the publications included in this dissertation collectively advance the field by tackling critical challenges in data curation, culture-aware evaluation, and model interpretation. The work delivers comprehensive insights into model behaviors while introducing practical approaches to enhance data efficiency. It establishes new benchmarks for evaluation within cultural contexts and identifies key VLM architectural limitations. These contributions provide a foundation for developing multimodal systems that are not only more robust and reliable but also more inclusive.

7.1 Open Problems and Future Directions

The ultimate goal of multimodal learning is to enable AI system to percept and reason about the world as humans do. Such system would dynamically process and adapt to diverse contexts while providing transparent reasoning for their decisions. Imagine tourists visiting historical landmarks being able to "relive" the past—seeing, hearing, and even experiencing sensory details like smells and tastes. We could foresee future multimodal systems that

provide personalized, culturally-informed experiences, transforming fields like cross-cultural understanding, education, health care and tourism. Based on our observations and insights, we outline key open problems and future directions in multimodal learning.

Cross-Cultural Multimodal Understanding

Figure 7.1: Culture-aware Multimodal AI

Building truly culturally-aware multimodal systems requires addressing three interconnected challenges:

a. We need large-scale, high-quality datasets that capture diverse cultural concepts and contexts, including temporal evolution of cultural practices. This requires interdisciplinary collaboration with experts from linguistics, cultural studies, and social sciences.

b. Develop representation learning techniques capable of handling cultural variations is critical in building more inclusive multimodal systems. This includes methods for disentangling culture-specific attributes from domain-general features, enabling models to distinguish universal concepts from culturally situated ones.

c. Beyond accuracy metrics for tasks like question answering, we need evaluation frameworks that assess cultural appropriateness and sensitivity across diverse demographic groups. This involves developing metrics that can identify subtle forms of cultural misrepresentation without reinforcing Western-centric perspectives as universal standards.

Interpretable Multimodal Reasoning

Current vision-language models struggle with real-world scenarios involving noisy or ambiguous visual information (Gurari et al., 2020), often producing misleading predictions without appropriate uncertainty signals. Although this dissertation has examined information loss in VLM connectors, the mechanisms by which visual information is selectively

preserved or discarded during text-vision interactions within language models remain poorly understood.

a. As observed in recent research (Groot and Valdenegro-Toro, 2024), vision-language models tend to display overconfidence even when predictions are incorrect. Understanding the sources of this overconfidence and developing effective calibration techniques remains an open challenge. This is particularly critical for safety-sensitive domains such as healthcare and security, where poorly calibrated confidence estimates could lead to severe consequences through misplaced trust in erroneous model outputs.

b. Developing measures to precisely quantify how each modality influences final predictions would significantly enhance model interpretability and transparency. Such attribution methods would enable researchers to identify modality-specific failure patterns in cross-modal reasoning, pinpoint when models inappropriately prioritize one modality over another, and ultimately guide more balanced architectural designs that appropriately weight information across various input channels.

Personalized and Adaptive Multimodal Learning

Personalized AI has tremendous potential to enhance daily life, influencing how people interact with environments and process emotional experiences.

a. While this dissertation has focused primarily on text and image inputs, temperature, taste, and tactile senses are equally important in human activities and emotions. Establishing datasets and evaluation benchmarks targeting true multimodality would bridge these representations with robotics and embodied AI, advancing multimodal learning applications.

b. Developing systems that detect cultural presuppositions in user queries without reinforcing stereotypes represents a critical challenge. Such systems must balance cultural awareness with ethical considerations around privacy and individual autonomy.

By addressing these open challenges, we can move beyond current limitations toward multimodal systems that are not just technically sophisticated but culturally informed, interpretable, and adaptable to diverse human needs and contexts.

Bibliography

The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Qwen2 technical report. 2024.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957, 2019.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai, 2024.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:36573–36589, 2022.

Mohammad Alsharid, Rasheed El-Bouri, Harshita Sharma, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. A course-focused dual curriculum for image captioning.

In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021. doi: 10.1109/ISBI48211.2021.9434055.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Sy21R9JAW.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. URL https://openaccess.thecvf.com/content_iccv_2015/papers/Antol_VQA_Visual_Question_ICCV_2015_paper.pdf.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023.

Viktar Atliha and Dmitrij Šešok. Text augmentation using bert for image captioning. *Applied Sciences*, 10(17):5978, 2020.

Hammad Ayyubi, Rahul Lokesh, Alireza Zareian, Bo Wu, and Shih-Fu Chang. Learning from children: Improving image-caption pretraining via curriculum. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023. doi: 10.18653/v1/2023.findings-acl.846. URL https://aclanthology.org/2023.findings-acl.846.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification, 2023.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442, 2016.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Emanuele Bugliarello, Laurent Sartran, Aishwarya Agrawal, Lisa Anne Hendricks, and Aida Nematzadeh. Measuring Progress in Fine-grained Vision-and-Language Understanding, May 2023. URL http://arxiv.org/abs/2305.07558.

Davide Caffagni, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Synthcap: Augmenting transformers with synthetic data for image captioning. In *International Conference on Image Analysis and Processing*, pages 112–123. Springer, 2023.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. Cultural Adaptation of Recipes. *Transactions of the Association for Computational Linguistics*, 12, January 2024a. ISSN 2307-387X. doi: 10.1162/tacl_a_00634. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00634/119279/Cultural-Adaptation-of-Recipes.

Yong Cao, Wenyan Li, Jiaang Li, Yifei Yuan, Antonia Karamolegkou, and Daniel Hershcovich. Exploring visual culture awareness in gpt-4v: A comprehensive probing. *arXiv preprint arXiv:2402.06015*, 2024b. URL https://arxiv.org/abs/2402.06015.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023.

Ting-Yun Chang and Robin Jia. Data curation alone can stabilize in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8123–8144, 2023.

Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19098–19107, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/papers/Chen_Grounding_Answers_for_Visual_Questions_Asked_by_Visually_Impaired_People_CVPR_2022_paper.pdf.

Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Chen_LION_Empowering_Multimodal_Large_Language_Model_with_Dual-Level_Visual_Knowledge_CVPR_2024_paper.pdf.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning (ICML)*, pages 1597–1607. PMLR, 2020.

Xin Chen, Hua Zhou, Yu Zhu, and Liang Diao. Chinesefoodnet: A large-scale image dataset for chinese food recognition. *arXiv preprint arXiv:1705.02743*, 2017.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Chen_InternVL_Scaling_up_Vision_Foundation_Models_and_Aligning_for_Generic_CVPR_2024_paper.pdf.

Zui Chen, Lei Cao, and Sam Madden. Lingua manga: A generic large language model centric system for data curation. *arXiv preprint arXiv:2306.11702*, 2023.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53, 2024.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems, 2024.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. URL https://arxiv.org/abs/2409.17146.

Ailin Deng, Zhirui Chen, and Bryan Hooi. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *arXiv preprint arXiv:2402.15300*, 2024.

Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.emnlp-main.98.

Xinzhi Dong, Chengjiang Long, Wenju Xu, and Chunxia Xiao. Dual graph convolutional networks with transformer and curriculum learning for image captioning. In *Proceedings of the 29th ACM International Conference on Multimedia*. Association for Computing Machinery, 2021. doi: 10.1145/3474085.3475439. URL https://doi.org/10.1145/3474085.3475439.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024. URL https://arxiv.org/abs/2401.08281.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3210. URL https://aclanthology.org/W16-3210.

E. Fix and J.L. Hodges. *Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties*. USAF School of Aviation Medicine, 1951. URL https://books.google.ch/books?id=4XwytAEACAAJ.

Philip Feldman Foulds, R James, and Shimei Pan. Ragged edges: The double-edged sword of retrieval-augmented chatbots. *arXiv preprint arXiv:2403.01193*, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Peng, Mengdan Tao, Ke Tang, Ye Wu, Xinyue Wang, Zekun Li, Xuhui Yang, Tao Lin, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

John C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. doi: 10.1007/BF02291478.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

Tobias Groot and Matias Valdenegro-Toro. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. *arXiv preprint arXiv:2405.02917*, 2024.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII*, page 417–434, Berlin, Heidelberg, 2020. Springer-Verlag. doi: 10.1007/978-3-030-58520-4_25. URL https://doi.org/10.1007/978-3-030-58520-4_25.

Hongkun Hao, Guoping Huang, Lemao Liu, Zhirui Zhang, Shuming Shi, and Rui Wang. Rethinking translation memory augmented neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023. doi: 10.18653/v1/2023.findings-acl.162. URL https://aclanthology.org/2023.findings-acl.162.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 9729–9738, 2020.

Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.

Cuong Hoang, Devendra Sachan, Prashant Mathur, Brian Thompson, and Marcello Federico. Improving robustness of retrieval augmented translation via shuffling of suggestions, 2022.

Cuong Hoang, Devendra Sachan, Prashant Mathur, Brian Thompson, and Marcello Federico. Improving retrieval augmented neural machine translation by controlling source and fuzzy-match interactions. In *Findings of the Association for Computational Linguistics: EACL 2023*, 2023. doi: 10.18653/v1/2023.findings-eacl.22. URL https://aclanthology.org/2023.findings-eacl.22.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Adam Gleave, Laurent Sifre, and Geoffrey Irving. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 30433–30448. Curran Associates, Inc., 2022.

Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7986–7994, 2018.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23369–23379, 2023.

Yuchen Huang, Zhiyuan Fan, Zhitao He, Sandeep Polisetty, Wenyan Li, and Yi R. Fung. Cultureclip: Empowering clip with cultural awareness through synthetic images and contextualized captions. *Conference on Language Modeling*, 2025.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.18653/v1/2023.emnlp-main.308. URL https://aclanthology.org/2023.emnlp-main.308.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. *International Conference on Machine Learning*, pages 4651–4664, 2021.

Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=99RpBVpLiX.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max W.F. Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. arXiv2405.01483, 2024.

M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015. doi: 10.1126/science.aaa8415. URL https://www.science.org/doi/abs/10.1126/science.aaa8415.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2022. URL https://proceedings.mlr.press/v162/kandpal22a.html.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2598339. URL https://doi.org/10.1109/TPAMI.2016.2598339.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1086. URL https://aclanthology.org/D14-1086/.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10258–10279, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.573. URL https://aclanthology.org/2024.emnlp-main.573/.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. 123(1). doi: 10.1007/s11263-016-0981-7. URL https://doi.org/10.1007/s11263-016-0981-7.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems 25*, 2012.

M Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. *Advances in neural information processing systems*, 23, 2010.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907, 2024.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577. URL https://aclanthology.org/2022.acl-long.577.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Bohao Li, Yuan Cheng, Zhe Gan, Jianfeng Liu, and Lijuan Wang. Seed-bench-2: Benchmarking multimodal large language models with text and image-based seed of thoughts. *arXiv preprint arXiv:2401.05831*, 2024a.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308, June 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Li_SEED-Bench_Benchmarking_Multimodal_Large_Language_Models_CVPR_2024_paper.pdf.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7241–7259. Association for Computational Linguistics, December 2022a. doi: 10.18653/v1/2022.emnlp-main.488. URL https://aclanthology.org/2022.emnlp-main.488/.

Jiaang Li, Yifei Yuan, Wenyan Li, Mohammad Aliannejadi, Daniel Hershcovich, Anders Søgaard, Ivan Vulić, Wenxuan Zhang, Paul Pu Liang, Yang Deng, and Serge Belongie. Ravenea: A benchmark for multimodal retrieval-augmented visual culture understanding. *arXiv preprint arXiv:2505.14462*, 2025a. URL https://arxiv.org/abs/2505.14462.

Jiaxuan Li, Duc Minh Vo, Akihiro Sugimoto, and Hideki Nakayama. Evcap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension. *arXiv preprint arXiv:2311.15879*, 2023a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023b. URL https://proceedings.mlr.press/v202/li23q.html.

Minghui Li, Yan Wan, and Jinping Gao. What drives the ethical acceptance of deep synthesis applications? a fuzzy set qualitative comparative analysis. *Computers in Human Behavior*, 133:107286, 2022c. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2022.107286.

Songtao Li and Hao Tang. Multimodal alignment and fusion: A survey. *arXiv preprint arXiv:2411.17040*, 2024. URL https://arxiv.org/abs/2411.17040.

Wenyan Li, Dong Li, Wanjing Li, Yuanjie Wang, Hai Jie, and Yiran Zhong. MAP: Low-data regime multimodal learning with adapter-based pre-training and prompting. In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 185–190, Gothenburg, Sweden, September 2023c. Association for Computational Linguistics. URL https://aclanthology.org/2023.clasp-1.19/.

Wenyan Li, Jiaang Li, Rita Ramos, Raphael Tang, and Desmond Elliott. Understanding retrieval robustness for retrieval-augmented image captioning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9285–9299, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.503. URL https://aclanthology.org/2024.acl-long.503/.

Wenyan Li, Jonas Lotz, Chen Qiu, and Desmond Elliott. The role of data curation in image captioning. In *Proceedings of the 18th Conference of the European Chapter of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 1074–1088, St. Julian's, Malta, March 2024d. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.65.

Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. FoodieQA: A multimodal dataset for fine-grained understanding of Chinese food culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19077–19095, Miami, Florida, USA, November 2024e. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1063. URL https://aclanthology.org/2024.emnlp-main.1063/.

Wenyan Li, Raphael Tang, Chengzu Li, Clemente Pasti, Vésteinn Snæbjarnarson, Caiqi Zhang, Ivan Vulić, Ryan Cotterell, and Anders Søgaard. Lost in embeddings: Information loss in vision-language models. *Under review*, 2025b.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022. URL https://papers.neurips.cc/paper_files/paper/2022/file/702f4db7543a7432431df588d57bc7c9-Paper-Conference.pdf.

Junyan Lin, Haoran Chen, Dawei Zhu, and Xiaoyu Shen. To preserve or to compress: An in-depth study of connector selection in multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, November 2024. doi: 10.18653/v1/2024.emnlp-main.325. URL https://aclanthology.org/2024.emnlp-main.325/.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'ar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. URL https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485. Association for Computational Linguistics, November 2021a. URL https://aclanthology.org/2021.emnlp-main.818.

Fenglin Liu, Shen Ge, and Xian Wu. Competence-based multimodal curriculum learning for medical report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3001–3012, Online, August 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.234. URL https://aclanthology.org/2021.acl-long.234.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F. Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries, 2025. URL https://arxiv.org/abs/2501.01282.

Xiao Liu, Yansong Feng, Jizhi Tang, Chengang Hu, and Dongyan Zhao. Counterfactual recipe generation: Exploring compositional generalization in a realistic scenario. *arXiv preprint arXiv:2210.11431*, 2022.

Zhixuan Liu, You Won Shin, Beverley-Claire Okogwu, Youngsik Yun, Lia Coleman, Peter Schaldenbrand, Jihie Kim, and Jean Oh. Towards equitable representation in text-to-image synthesis models with the cross-cultural understanding benchmark (ccub) dataset. *ArXiv*, abs/2301.12073, 2023b. URL https://api.semanticscholar.org/CorpusID:256389945.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556.

Zheng Ma, Mianzhi Pan, Wenhan Wu, Kanzhi Cheng, Jianbing Zhang, Shujian Huang, and Jiajun Chen. Food-500 Cap: A Fine-Grained Food Caption Benchmark for Evaluating Vision-Language Models, August 2023. URL http://arxiv.org/abs/2308.03151.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*, 2023.

W. Min, B. K. Bao, S. Mei, Y. Zhu, Y. Rui, and S. Jiang. You are what you eat: Exploring rich recipe information for cross-region food analysis. *IEEE Transactions on Multimedia*, 20(4):950–964, 2018.

David Orlando Romero Mogrovejo, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesus-German Ortiz-Barajas, Emilio Villa Cueva, Jinheon Baek, Soyeong Jeong, et al. Cvqa: Culturally-diverse multilingual visual question answering benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.329. URL https://aclanthology.org/2024.emnlp-main.329/.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y Ng, et al. Multimodal deep learning. In *ICML*, volume 11, pages 689–696, 2011.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36:22047–22069, 2023.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang, editors, *ICML*, volume 139, pages 8162–8171. PMLR, 2021.

Malvina Nikandrou, Georgios Pantazopoulos, Nikolas Vitsakis, Ioannis Konstas, and Alessandro Suglia. Crope: Evaluating in-context adaptation of vision and language models to culture-specific concepts. *arXiv preprint arXiv:2410.15453*, 2024.

Asmaa AE Osman, Mohamed A Wahby Shalaby, Mona M Soliman, and Khaled M Elsayed. A survey on attention-based models for image captioning. *International Journal of Advanced Computer Science and Applications*, 14(2), 2023.

Shramay Palta and Rachel Rudinger. FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models. In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.631. URL https://aclanthology.org/2023.findings-acl.631.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.567. URL https://aclanthology.org/2022.acl-long.567.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.196. URL https://aclanthology.org/2022.findings-acl.196.

Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. doi: 10.18653/v1/2022.emnlp-main.731. URL https://aclanthology.org/2022.emnlp-main.731.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger,

and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140): 1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pages 18–34, 2024.

Vikram V. Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron B. Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition. In *NeurIPS Datasets and Benchmarks*, 2023.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ramesh21a.html.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022.

Rita Ramos, Desmond Elliott, and Bruno Martins. Retrieval-augmented image captioning. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3666–3681, Dubrovnik, Croatia, May 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.266. URL https://aclanthology.org/2023.eacl-main.266.

Rita Ramos, Bruno Martins, and Desmond Elliott. LMCap: Few-shot multilingual image captioning by retrieval augmented language model prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, July 2023b. URL https://aclanthology.org/2023.findings-acl.104.

Rita Ramos, Bruno Martins, and Desmond Elliott. Smallcap: Lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23628–23638, 2023c.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Anna Rogers. Changing the world by changing the data. In *Annual Meeting of the Association for Computational Linguistics*, 2021. URL https://api.semanticscholar.org/CorpusID:235248305.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. Cvqa: Culturally-diverse multilingual visual question answering benchmark, 2024.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, 2022.

Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Retrieval-augmented transformer for image captioning. In *Proceedings of the 19th International Conference on Content-Based Multimedia Indexing*, CBMI '22, 2022. ISBN 9781450397209. doi: 10.1145/3549555.3549585. URL https://doi.org/10.1145/3549555.3549585.

Simon Schrodi, David T Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language representation learning. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024. URL https://openreview.net/forum?id=7QwFMLzQHH.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint*, 2021. URL https://doi.org/10.48550/arXiv.2111.02114.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: an open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238/.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arxiv:2305.17493*, 2023.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3463257. URL https://doi.org/10.1145/3404835.3463257.

Barry E Stein and M Alex Meredith. *The merging of the senses*. MIT press, 1993.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Sun_Generative_Multimodal_Models_are_In-Context_Learners_CVPR_2024_paper.pdf.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 2017.

Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning (ICML)*, pages 6105–6114. PMLR, 2019.

Raphael Tang, Crystina Zhang, Lixinyu Xu, Yao Lu, Wenyan Li, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. Words worth a thousand pictures: Measuring and understanding perceptual variability in text-to-image generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5441–5454, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.311. URL https://aclanthology.org/2024.emnlp-main.311/.

David Thiel. Investigation finds ai image generation models trained on child abuse, Dec 2023. URL https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248, June 2022.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann Lecun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings - 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, 2024. doi: 10.1109/CVPR52733.2024.00914.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

Alan Turing. Computing machinery and intelligence. *Mind*, 59(236):433–60, 1950. doi: 10.1093/mind/lix.236.433.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Emiel van Miltenburg and Desmond Elliott. Room for improvement in automatic image description: an error analysis. *CoRR*, abs/1704.04198, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2015.html#VedantamZP15.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4808. URL https://aclanthology.org/W19-4808.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. Caltech-ucsd birds 200. Technical report, California Institute of Technology, 2011. URL https://authors.library.caltech.edu/records/cvm3y-5hh21.

Shenzhi Wang and Yaowei Zheng. Llama3-8b-chinese-chat (revision 6622a23), 2024. URL https://huggingface.co/shenzhi-wang/Llama3-8B-Chinese-Chat.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023a.

Xiao Wang, Ibrahim Alabdulmohsin, Daniel Salz, Zhe Li, Keran Rong, and Xiaohua Zhai. Scaling pre-training to one hundred billion data for vision language models. *arXiv preprint arXiv:2502.07617*, 2025.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. " my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. *arXiv preprint arXiv:2402.14499*, 2024.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. Learning to filter context for retrieval-augmented generation, 2023b.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*, 2022.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, Enrico Santus, Fariz Ikhwantri, Garry Kuwanto, Hanyang Zhao, Haryo Akbarianto Wibowo, Holy Lovenia, Jan Christian Blaise Cruz, Jan Wira Gotama Putra, Junho Myung, Lucky Susanto, Maria Angelica Riera Machin, Marina Zhukova, Michael Anugraha, Muhammad Farid Adilazuarda, Natasha Christabelle Santosa, Peerat Limkonchotiwat, Raj Dabre, Rio Alexander Audino, Samuel Cahyawijaya, Shi-Xiong Zhang, Stephanie Yulia Salim, Yi Zhou, Yinxuan Gui, David Ifeoluwa Adelani, En-Shiun Annie Lee, Shogo Okada, Ayu Purwarianti, Alham Fikri Aji, Taro Watanabe, Derry Tanti Wijaya, Alice Oh, and Chong-Wah Ngo. WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume*

1: Long Papers), pages 3242–3264, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.167. URL https://aclanthology.org/2025.naacl-long.167/.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Junru Wu, Yi Liang, Hassan Akbari, Zhangyang Wang, Cong Yu, et al. Scaling multimodal pre-training via cross-modality gradient harmonization. *Advances in Neural Information Processing Systems*, 35:36161–36173, 2022.

Changrong Xiao, Sean Xin Xu, and Kunpeng Zhang. Multimodal data augmentation for image captioning using diffusion models. *arXiv preprint arXiv:2305.01855*, 2023.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes, September 2018. URL http://arxiv.org/abs/1809.00812.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024.

Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, et al. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. *arXiv preprint arXiv:2302.04858*, 2023.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. ICML'23. JMLR.org, 2023.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*, 2023.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.

Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, Sam Dodge, Keen You, Zhen Yang, Aleksei Timofeev, Mingze Xu, Hong-You Chen, Jean-Philippe Fauconnier, Zhengfeng Lai, Haoxuan You, Zirui Wang, Afshin Dehghan, Peter Grasch, and Yinfei Yang. MM1.5: Methods, analysis & insights from multimodal LLM fine-tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=HVtu26XDAA.

Hongkuan Zhang, Saku Sugawara, Akiko Aizawa, Lei Zhou, Ryohei Sasano, and Koichi Takeda. Cross-modal similarity-based curriculum learning for image captioning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022a. doi: 10.18653/v1/2022.emnlp-main.516. URL https://aclanthology.org/2022.emnlp-main.516.

Na Zhang and Guansheng Ma. Nutritional characteristics and health effects of regional cuisines in china. *Journal of Ethnic Foods*, 7, 12 2020. doi: 10.1186/s42779-020-0045-z.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5014–5022, 2016.

Peng Zhang, Can Li, Liang Qiao, Zhanzhan Cheng, Shiliang Pu, Yi Niu, and Fei Wu. Vsr: a unified framework for document layout analysis combining vision, semantics and relations. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16*, pages 115–130. Springer, 2021.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022b.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/5c7024041be305c94d7311cfcc53d93e-Paper-Conference.pdf.

Li Zhou, Taelin Karidi, Nicolas Garneau, Yong Cao, Wanlong Liu, Wenyu Chen, and Daniel Hershcovich. Does mapo tofu contain coffee? probing llms for food-related cultural knowledge. *arXiv preprint arXiv:2404.06833*, 2024.

Li Zhou, Lutong Yu, Dongchu Xie, Shaohuan Cheng, Wenyan Li, and Haizhou Li. Hanfubench: A multimodal benchmark on cross-temporal cultural understanding and transcreation. *arXiv preprint arXiv:2506.01565*, 2025. URL https://arxiv.org/abs/2506.01565.

Yucheng Zhou and Guodong Long. Style-aware contrastive learning for multi-style image captioning. In *Findings of the Association for Computational Linguistics: EACL 2023*, 2023. doi: 10.18653/v1/2023.findings-eacl.169. URL https://aclanthology.org/2023.findings-eacl.169.