UniRAG: Universal Retrieval Augmentation for Multi-Modal Large Language Models

Sahel Sharifymoghaddam*, **Shivani Upadhyay***, **Wenhu Chen**, **Jimmy Lin**

Department of Computer Science
University of Waterloo

{sahel.sharifymoghaddam, sjupadhyay, wenhu.chen, jimmylin}@uwaterloo.ca

Abstract

Recently, Multi-Modal (MM) Large Language Models (LLMs) have unlocked many complex use-cases that require MM understanding (e.g., image captioning or visual question answering) and MM generation (e.g., text-guided image generation or editing) capabilities. To further improve the output fidelity of MM-LLMs we introduce UniRAG, a plug-and-play technique that adds relevant retrieved information to prompts as few-shot examples during inference. Unlike the common belief that Retrieval Augmentation (RA) mainly improves generation or understanding of uncommon entities, our evaluation results on the MSCOCO dataset with common entities show that both proprietary models like GPT-4o and Gemini-Pro and smaller open-source models like LLaVA, LaVIT, and Emu2 significantly enhance their generation quality when their input prompts are augmented with relevant information retrieved by MM retrievers like UniIR models. All the code required for reproducing our results will be released upon acceptance.

保真度，逼真度

1 Introduction

Recent advancements in Multi-Modal Large Language Models (MM-LLMs), encompassing both proprietary models like GPT-4o (OpenAI et al., 2024), Gemini-Pro (Team et al., 2024), DALL-E (Ramesh et al., 2021) and Parti (Yu et al., 2022) and open-source models such as LLaVA (Liu et al., 2023c,b), LaVIT (Jin et al., 2023) and Emu2 (Sun et al., 2023), are instrumental in bridging the gap between different modalities. These models have demonstrated remarkable human-like efficacy and achieved state-of-the-art effectiveness in various benchmark assessments (Li and Lu, 2024).

效力，功效

However, like LLMs, MM-LLMs often struggle to produce accurate results on lesser-known or recent topics (Huang et al., 2024; Bai et al., 2024). This limitation occurs because models generate responses solely based on their training data, lacking access to external information during inference. Given the rapidly changing nature of the world, it is unrealistic for models to correctly address every query using only pre-trained knowledge.

As a workaround, techniques like Retrieval Augmented Generation (RAG) have gained significant attention in recent years (Gao et al., 2023). In-context RAG (Ram et al., 2023) particularly relies on the learning capabilities of models to incorporate the latest knowledge and guide them in generating more relevant results during inference. Recent studies have applied RAG to multi-modal tasks, including visual question answering, text-guided image generation and editing, and image captioning (Yasunaga et al., 2022; Lin and Byrne, 2022; Chen et al., 2022a; Yu et al., 2023; Hu et al., 2023). However, these approaches mainly focus on training models with large image-text datasets, leaving the potential for plug-and-play in-context RAG in MM applications largely unexplored. As black-box MM-LLMs become more prevalent, addressing this gap is essential for their reliable adoption in applications that cannot tolerate hallucinations.

多模态的一些任务

This paper introduces the plug-and-play UniRAG technique, designed to enhance the fidelity of model outputs. UniRAG integrates RA with MM-LLMs to guide the generation process using relevant in-context information. By incorporating interleaved image-text pairs as few-shot examples during inference, UniRAG serves as a model-agnostic approach that effectively teaches out-of-the-box MM-LLMs about various task intents and uncommon entities. Our evaluation in this work primarily focuses on image captioning (image-to-text) and image generation (text-to-image) tasks.

交叉存取的

UniRAG adopts a two-stage retrieval and generation approach. For retrieval we use the UniIR (Wei et al., 2023) retriever, which is trained with diverse MM datasets to retrieve heterogeneous out-

各种各样的

* Equal contribution.

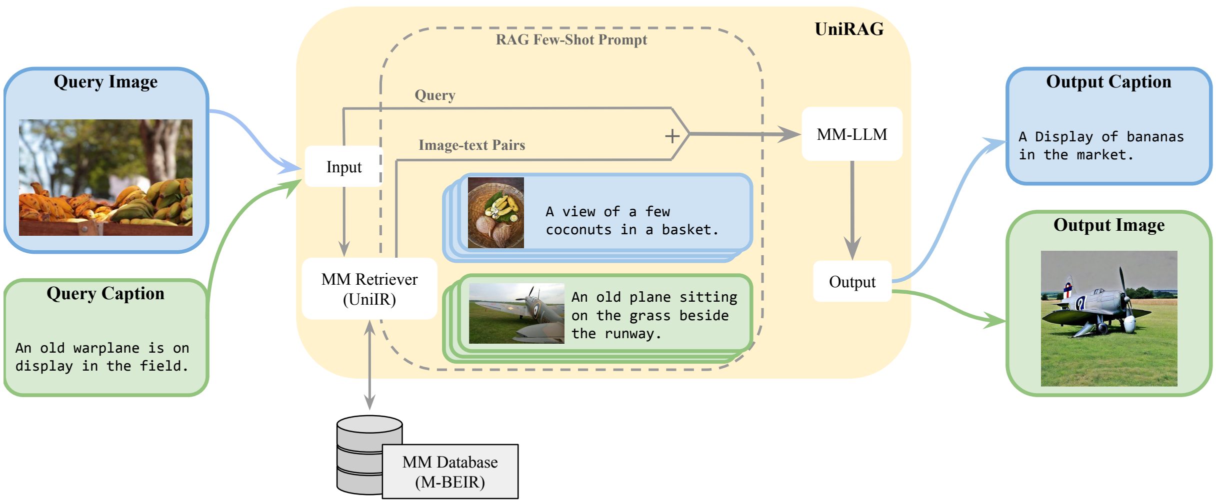


Figure 1: An Overview of the UniRAG technique with image captioning (blue) and image generation (green) tasks. UniRAG retrieves relevant image-text pairs and adds them as few-shot examples to the MM-LLM's input prompt.

puts in both text and image modalities. We mainly adopt UniIR's CLIP Score Fusion and BLIP Feature Fusion models which are instruction-tuned using CLIP (Radford et al., 2021) and BLIP-2 (Li et al., 2022), respectively. In the generation stage, we use a variety of off-the-shelf proprietary and open-source MM-LLMs models. In particular, we employ LLaVA, GPT-4o and Gemini-Pro for image captioning and LaVIT and Emu2 for image generation. We guide these MM-LLMs using zero-shot and few-shot prompting techniques to demonstrate their enhanced generation capabilities due to access to relevant examples during inference.

现成的

For our evaluations, we mainly utilize the MSCOCO (Chen et al., 2015) dataset from the M-BEIR (Wei et al., 2023) benchmark. We assess caption and image generation using M-BEIR's image-to-text and text-to-image tasks, respectively. By combining RA with MM-LLMs, we achieve significant improvements over baseline effectiveness: an average increase of about 9 percentage points in SPICE (Anderson et al., 2016) for image captioning and a reduction of 25 units in Fréchet Inception Distance (FID) (Heusel et al., 2017) for image generation. Additional evaluations using M-BEIR's Fashion200k dataset demonstrate that UniRAG is even more effective in domain-specific applications. To summarize, our main contributions in this paper are as follows:

- Introducing the plug-and-play UniRAG technique, which combines MM retrievers with MM-LLMs using in-context RAG.
- Extensive evaluation of UniRAG's effectiveness for image captioning and image generation tasks using five MM-LLMs including both open-source and proprietary models.
- Evaluating UniRAG on the Fashion200k dataset to showcase its effectiveness in domain-specific applications.

2 Related Work

Retrieval Augmentation with Generative Models: Retrieval Augmentation (RAG) techniques have significantly improved the effectiveness of LLM generation by incorporating external information, leading to higher-quality results (Gao et al., 2023). While some researchers modify LLM architectures or fine-tune them for conditioning on retrieved information (Guu et al., 2020; Borgeaud et al., 2022), others explore in-context RAG (Ram et al., 2023; Dehghan et al., 2024), which incorporates relevant data directly into prompts. Recent studies have also examined the integration of RAG with chain-of-thought (Wei et al., 2022; Wang et al., 2023b) and other prompting techniques (Wang et al., 2023a; Li et al., 2024); however, most of these works mainly augment generation with text.

Multi-Modal Models and Retrieval Augmentation (Vision-Language): Recently, a number of models have been developed that integrate vision and language understanding (Ramesh et al., 2021; Radford et al., 2021; Yu et al., 2022; Li et al., 2022; Liu et al., 2023c,b; Jin et al., 2023; Sun et al., 2023; Zhai et al., 2023; OpenAI et al., 2024;

Team et al., 2024). These models typically use modality-specific encoders trained on large image-text datasets to form a unified feature space. More details on the MM-LLMs selected for evaluation can be found in Section 3.2.

Similar to traditional LLMs, MM-LLMs also struggle with generating content about lesser-known entities or uncommon combinations of common ones (Chen et al., 2022b; Blattmann et al., 2022). By leveraging retrieval, MM-LLMs can be guided to produce more accurate and high-fidelity outputs. UniIR (Wei et al., 2023) introduces MM retriever models that are instruction-tuned with diverse MM datasets to perform heterogeneous retrieval tasks involving both text and images. RA-CM3 (Yasunaga et al., 2022), MuRAG (Chen et al., 2022a), CM3Leon (Yu et al., 2023) and REVEAL (Hu et al., 2023) employ a retriever-generator workflow for RAG in various MM tasks. However, they all require training generator and/or retriever models on extensive MM datasets.

现有的这些方法都需要训练

In contrast, UniRAG incorporates an MM retriever during inference, facilitating in-context RAG for black-box LLMs. The work by Liu et al. (2023a), which presents an RA chain-of-thought method, is closest to ours. However, unlike UniRAG, it focuses on improving the effectiveness of MM-LLMs for visual question answering.

3 Methodology

UniRAG adopts a two-stage retrieval and generation workflow that is explained below:

3.1 Retrieval

The retrieval stage is where for a given query its top k most relevant candidates are extracted from an MM database. In this stage query and candidate modalities can be any combination of $\{text, image\} \rightarrow \{text, image\}$. Specifically, for an image-to-text task such as caption generation, the query is an image, and retrieved candidates are captions; while for a text-to-image task (e.g., image generation), the query is a caption and retrieved candidates are images.

两种任务，图像解释和图像生成

To use image-text pairs as in-context examples in the next stage, we need to convert the retrieved single-modality candidates into pairs. This involves treating each retrieved candidate as a query to find its complementary candidate based on two criteria: a) it must have the opposite modality (text for an image candidate and vice versa) and b) it must differ from the original query to prevent revealing the ground-truth answer. The retriever models used for our experiments are further explained in Section 4

3.2 Generation

The generation stage is where the MM-LLM generates the required output in zero-shot (Kojima et al., 2023) or few-shot (Brown et al., 2020) settings. Similar to the retrieval stage, in the zero-shot setting (baseline 1), the query is only present in the input prompt; while in the few-shot setting, in-context examples are additionally included in the prompt. In this setting, examples are obtained using one of the following methods: 1) random selection of queries and their ground truth candidates from the dataset (baseline 2); 2) in-context RAG with retrieved candidate pairs from the previous stage. We leverage various MM-LLMs for caption and image generation tasks, as detailed in Section 4.

4 Experimental Setup

4.1 Selected Models

We use UniIR's CLIP Score Fusion (CLIP-SF) and BLIP Feature Fusion (BLIP-FF) models as MM retrievers. Experimental results from the original paper indicate that these instruction-tuned UniIR models significantly outperform their pre-trained CLIP and BLIP2 baselines across various tasks and configurations, including image and caption retrieval tasks (Wei et al., 2023).

To generate appropriate text captions for images, we use the LLaVA (13B) (Liu et al., 2023b), Gemini-Pro (Team et al., 2024), and GPT-4o (OpenAI et al., 2024) models. Table 1 presents sample captions generated by these models in zero-shot, random few-shot, and RAG few-shot settings (retrieved by CLIP-SF and BLIP-FF).

LLaVA is an open-source MM model based on Vicuna (Chiang et al., 2023), utilizing CLIP as its visual encoder. It is fine-tuned on MM instruction-following data generated by GPT-4 (Liu et al., 2023c). LLaVA's ability to understand and follow human intent makes it well-suited for RAG caption generation. Figures 2- 5 in Appendix A show detailed prompts used for caption generation.

Unlike image captioning, which only requires MM understanding, image generation necessitates models with both MM understanding and generation capabilities. To meet this need, we selected LaVIT and Emu2-Gen from a survey of MM-LLMs (Zhang et al., 2024).

	Image Query	Top retrieved image-caption pair ($k = 1$)		
	Zeroshot	CLIP-SF	BLIP-FF	Random
Prompt				
	-	Black and white photograph of two men on motorcycles.	A man riding on the back of a motorcycle down a highway.	A pair of red scissors sitting on a newspaper.
LLaVA	Motorcycle rider on the road, with luggage attached to the back.	Black and white photograph of a man on a motorcycle.	A man riding on the back of a motorcycle down a highway.	A man is sitting on a motorcycle, and there are two and a half hot-dogs on paper plates on a counter.
Gemini-P	Biker and passenger riding down the road together.	Black and white photograph of a man and woman riding on a motorcycle.	Two people riding on a motorcycle through a parking lot.	Biker and his passenger riding down the road.
GPT-4o	two people on a motorcycle waiting at a traffic light.	black and white photograph of two men on a motorcycle at an intersection.	two people riding on a motorcycle through an intersection.	two people are sharing a large cruiser motorcycle at a stoplight.

Table 1: Sample caption generation with LLaVA, Gemini-Pro, and GPT-4o models in zero-shot and one-shot settings. The "Prompt" row shows the zero-shot image query as well as retrieved image-caption pairs from CLIP-SF, BLIP-FF and random selection that are included in the prompt as in-context examples.

LaVIT is an open-source model based on LLaMA-7B (Touvron et al., 2023) that employs a visual tokenizer to convert images into discrete visual tokens. This allows it to process both text and image inputs using a unified objective for next image/text token prediction. The authors report a Frechet Inception Distance (FID) of 7.4 (Heusel et al., 2017) on the MSCOCO dataset, indicating its strong performance in image generation tasks.

Emu2-Gen is another open-source model, finetuned from Emu2 for image generation. Like LaVIT, it uses next token prediction across all modalities and incorporates a variant of CLIP as its visual encoder along with LLaMA-33B as its backbone. It can accept interleaved inputs of images, text, and video, and features a visual decoder for generating images. While its zero-shot performance on various MM tasks is lower than that of Emu (Dai et al., 2023), Flamingo (Alayrac et al., 2022), and others, it excels in few-shot settings, making it suitable for image generation with RAG.

Table 2 displays visual examples of image generation using LaVIT and Emu2-Gen in zero-shot, random few-shot, and RAG few-shot settings (using CLIP-SF as the retriever).

4.2 Dataset

We evaluate UniRAG on image-to-text and text-to-image tasks from the M-BEIR (Wei et al., 2023) benchmark using its MSCOCO and Fashion200k datasets. To ensure generic retrieval, we retrieve from M-BEIR's global candidate pool, which contains over 5.5 million candidates from 10 different datasets, instead of using dataset-specific pools.

The MSCOCO test set includes about 25k captions for 5k unique images of common objects. For caption generation, we use all 5k images as queries. However, due to the lengthy image generation time (over 12 seconds per image), we randomly sample one caption query for each image. To maintain consistency, all image generation runs use this same sample set. We analyze the effects of this sampling in detail in section 5.2.1. To demonstrate UniRAG's effectiveness in domain-specific tasks, we also evaluate it on the test set of Fashion200k dataset from M-BEIR. This set includes about 1.7k caption and 4.9k image queries, all focused on the fashion domain.

4.3 Configuration Details

We use the LLaVA, Gemini-Pro, and GPT-4o models for generating captions, and LaVIT and Emu2-Gen for image generation. For both tasks, we experiment with adding $k \in \{0, 1, 5\}$ retrieved image-caption pairs as in-context examples, where $k = 0$ establishes the baseline effectiveness of generator models without random or RAG in-context examples. Since LLaVA does not allow multiple images

	Caption Query	**Top retrieved image-caption pair** ($k = 1$)		
	Zeroshot	CLIP-SF	BLIP-FF	Random
Prompt	-			
	A large jetliner sitting on top of an airport runway.	Black and white jet airliner on an airport runway.	A large plane is parked on the run way.	Two teenage boys are playing a game of frisbee together.
LaVIT				
Emu2-G				

Table 2: Sample image generation with LaVIT, and Emu2-Gen models in zero-shot, and one-shot settings. The "Prompt" row shows the zero-shot caption query as well as retrieved image-caption pairs from CLIP-SF, BLIP-FF and random selection that are included in the prompt as in-context examples.

LlaVa只能接收一张图像，GPT-4o能接收多张图像，所以图像的处理方式不同哦

as input, we vertically merge all few-shot candidate images (and the query image for caption generation) into a single image. The final merged image's width matches the widest image, while its height is the total of all individual heights. Figure 6 shows a sample merged image with $k = 1$ examples. Each image is then labeled as $[n]$ in the prompt, where n indicates its position in the final merged image. For all other generator models, the in-context examples are provided in the following interleaved format: $\{\text{image}_1, \text{text}_1, \ldots, \text{image}_k, \text{text}_k\}$.

We use LLaVA's default configuration [1] for all its inferences, with a batch size of four for zero-shot and two for few-shot generation. Zero-shot caption generation on a single NVIDIA RTX A6000 GPU averages around 5 seconds, while few-shot takes 2-3 times longer. For Gemini-Pro and GPT-4o, we utilize Vertex AI's *generate content* and OpenAI's *chat completion* APIs, respectively, setting their `max_new_token` parameter to 400. More details on their cost estimates can be found in Appendix B. For LaVIT, we follow the sample configuration provided [2], generating a 1024×1024 image in about 12 seconds on a single NVIDIA RTX A6000 GPU. Similarly, we use Emu2-Gen's default settings [3]. Since it has 37 billion parameters, we run its "bf16" variant on an NVIDIA H100 SXM from the Lambda GPU provider. Image generation with this setting takes about 8 seconds for zero-shot and one-shot, and 13 seconds for five-shot. We encountered CUDA out-of-memory errors for 36 out of 5k captions during five-shot image generation with Emu2-Gen, so we limited the prompt to the first four pairs in those cases.

For each task, we report the most commonly used metrics: For caption generation, we include n-gram based metrics, BLEU(1-4) (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), and ROUGE (Lin, 2004), as well as the scene-graph-based metric, SPICE. For image generation, we report FID to measure image fidelity, CLIP Score (Hessel et al., 2021) for alignment with the caption prompt, and Inspection Score (IS) (Salimans et al., 2016) for overall quality. Appendix D explains these metrics in more detail.

[1] https://github.com/huggingface/transformers/blob/main/src/transformers/models/llava/configuration_llava.py

[2] https://github.com/jy0205/LaVIT/blob/main/LaVIT/text2image_synthesis.ipynb

[3] https://github.com/baaivision/Emu/tree/main/Emu2#emu2-gen

	k	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE	SPICE
(1a) CLIP-SF	1	88.5	84.2	81.5	79.9	215.3	83.2	37.6
(1b) BLIP-FF	1	89.4	84.6	81.5	79.5	218.2	83.9	37.6
(2) LLaVA	0	58.0	39.9	27.0	18.0	64.3	42.2	17.1
(2a) CLIP-SF +	1	80.1	70.7	63.9	59.0	162.9	70.3	31.1
LLaVA	5	75.1	61.7	50.7	42.4	128.9	60.6	24.8
(2b) BLIP-FF +	1	**81.0**	**71.5**	**64.5**	**59.4**	**166.9**	**71.1**	**31.1**
LLaVA	5	75.5	61.9	50.7	42.2	129.0	60.6	25.0
(2c) Random +	1	62.6	45.7	32.7	23.1	79.2	47.5	16.9
LLaVA	5	46.1	27.6	16.1	9.9	27.9	36.1	7.0
(3) Gemini-P	0	45.1	29.5	19.4	12.8	43.8	30.4	11.8
(3a) CLIP-SF +	1	64.6	47.4	34.9	26.1	83.3	47.9	20.2
Gemini-P	5	69.4	52.4	38.9	28.6	**96.4**	**51.6**	**21.7**
(3b) BLIP-FF +	1	64.3	47.1	34.4	25.4	83.4	48.2	20.2
Gemini-P	5	**69.9**	**53.1**	**39.6**	**29.4**	94.4	49.9	21.0
(3c) Random +	1	64.0	45.9	32.1	22.2	79.4	46.6	18.3
Gemini-P	5	69.0	51.1	36.8	25.9	92.4	50.7	20.5
(4) GPT-4o	0	59.6	40.1	26.4	17.0	68.3	44.9	18.5
(4a) CLIP-SF +	1	56.6	39.0	26.8	18.7	63.3	45.5	17.7
GPT-4o	5	65.4	48.4	35.2	25.7	87.5	50.3	20.5
(4b) BLIP-FF +	1	57.9	40.3	28.0	19.8	67.1	46.6	18.7
GPT-4o	5	**66.4**	**49.2**	**36.1**	**26.3**	**90.4**	**51.0**	**21.1**
(4c) Random +	1	60.2	42.0	28.6	19.2	72.5	46.5	18.7
GPT-4o	5	55.8	38.2	26.1	17.9	61.1	42.2	14.5

Table 3: Caption generation with LLaVA, Gemini-Pro and GPT-4o models on the MSCOCO's dataset using UniIR's CLIP-SF and BLIP-FF as retrievers. Column $k \in \{0, 1, 5\}$ shows the number of few-shot examples in each run.

	k	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE	SPICE
(1a) CLIP-SF	1	35.9	18.8	13.2	10.8	90.2	36.6	21.7
(1b) BLIP-FF	1	41.8	25.3	19.2	16.6	130.4	41.7	26.5
(2) LLaVA	0	9.7	2.5	0.5	0.0	6.7	14.0	10.0
(2a) CLIP-SF +	1	32.9	16.7	11.4	9.0	75.1	33.6	19.9
LLaVA	5	34.6	15.8	9.3	6.5	63.8	35.2	20.2
(2b) BLIP-FF +	1	38.9	**22.8**	**16.7**	**14.0**	**111.1**	38.5	**24.1**
LLaVA	5	**39.2**	20.3	13.2	10.0	86.7	**39.2**	23.1
(2c) Random +	1	22.8	6.3	1.6	0.6	22.3	23.4	12.2
LLaVA	5	11.9	2.6	0.8	0.3	8.0	12.6	4.8

Table 4: Caption generation with LLaVA on the Fashion200k using UniIR's CLIP-SF and BLIP-FF as retrievers. Column $k \in \{0, 1, 5\}$ represents the number of few-shot examples for each experiment.

5 Evaluation Results and Analysis

Tables 3- 6 present results for caption and image generation tasks on MSCOCO and Fashion200k datasets. Rows 1* from each table measure the effectiveness of CLIP-SF and BLIP-FF retrievers. They reflect how well the generator models perform if they return the image or caption from the first retrieved example as is. To ensure an equal number of images and captions for evaluation, we report retrievers' performance using only their top retrieved image-text pair ($k = 1$) for each query.

CLIP-SF and BLIP-FF confuse candidate modalities for up to 7% and 25% of MSCOCO queries, respectively, meaning they may retrieve an image for an image-to-text or a text for a text-to-image task. In our evaluation, when a retrieved candidate has the wrong modality, we use its complement with the correct modality. Unlike the original UniIR results, which indicated CLIP-SF was more effective for MSCOCO retrieval, this adjustment allows both retrievers to perform similarly on this dataset. For Fashion200k, our findings confirm UniIR's results, showing BLIP-FF is more effective.

The remaining rows in each table compare zero-shot ($k = 0$) and few-shot ($k > 0$) results of generator models using CLIP-SF (*a), BLIP-FF (*b), and random selection (*c) as retrievers for few-shot generations. The performance of CLIP-SF and BLIP-FF retrievers in RAG applications aligns with their standalone effectiveness (comparing rows *a and *b in each table). On MSCOCO, both retrievers continue to perform similarly, while BLIP-FF consistently outperforms CLIP-SF on Fashion200k.

5.1 Caption Generation

Table 3 shows the caption generation results for the LLaVA, Gemini-Pro and GPT-4o generator models on the MSCOCO dataset. While several metrics are reported in this table as a reference, we focus on SPICE since all other metrics are primarily sensitive to n-gram overlaps (Anderson et al., 2016).

As reported in Table 3, augmenting relevant image-caption pairs ($k > 0$) retrieved from the previous stage, boosts the effectiveness of MM-LLMs regardless of their baseline effectiveness, the number of in-context examples and the choice of retriever model. The baseline effectiveness ($k = 0$) of LLaVA exceeds that of Gemini-Pro and GPT-4o models (comparing rows 2-4). Even though all models benefit from RA, this gap further widens when only a single retrieved example ($k = 1$) is added to the prompts, making one-shot BLIP-FF + LLaVA the most effective setting and one-shot CLIP-SF + LLaVA a very close runner up. Further increasing the number of retrieved examples from one to five helps the Gemini-P and GPT-4o models. However, LLaVA with five retrieved examples is less effective than LLaVA with a single example.

For both LLaVA and GPT-4, adding a single randomly selected example to the prompt slightly improves the generation quality compared to zero-shot generation. However, increasing the number of randomly selected examples to five degrades the generators' effectiveness to below their zero-shot baselines. In contrast, for Gemini-Pro, random few-shot examples significantly enhance the quality of generated captions, although they are still less effective than few-shot retrieved examples. These trends show that UniRAG's model-agnostic effectiveness comes from only including the most relevant examples in-context.

Table 4 shows the caption generation results for the Fashion200k dataset. To reduce the cost of API calls, we only report results for the LLaVA model in this experiment. LLaVA performs poorly in zero-shot scenarios due to its limited knowledge of the domain-specific Fashion200k dataset. While randomly selected few-shot examples can help LLaVA grasp the task intent and the required level of detail for describing fashion products, the real improvement occurs when relevant examples are included in the prompts. Because many entities in the Fashion200k dataset are quite similar, the generator often repeats the in-context caption of a similar image for its query image, leading to effectiveness that is close to that of the baseline retriever models. Please see Appendix E for sample visualizations.

5.2 Image Generation

	k	FID ↓	CLIP Score ↑	IS ↑ (SD)
(1a) CLIP-SF	1	5.88	29.99	22.19 (1.29)
(1b) BLIP-FF	1	7.01	29.97	22.77 (1.02)
(2) LaVIT	0	64.80	26.27	16.38 (1.27)
(2a) CLIP-SF + LaVIT	1	23.39	30.42	24.11 (1.09)
	5	24.14	31.09	24.52 (1.08)
(2b) BLIP-FF + LaVIT	1	**23.23**	30.44	24.75 (1.25)
	5	23.95	31.07	**25.23 (1.10)**
(2c) Random + LaVIT	1	26.90	22.43	19.31 (0.90)
	5	26.46	**31.13**	22.89 (1.35)
(3) Emu2-G	0	41.66	29.54	21.43 (0.73)
(3a) CLIP-SF + Emu2-G	1	**30.99**	30.02	**23.51 (0.98)**
	5	32.51	29.94	22.01 (1.02)
(3b) BLIP-FF + Emu2-G	1	31.37	**30.04**	23.27 (0.75)
	5	33.17	29.96	22.63 (1.84)
(3c) Random + Emu2-G	1	38.49	23.62	16.86 (1.03)
	5	53.92	23.74	11.51 (0.32)

Table 5: Image generation with LaVIT and Emu2-G on the MSCOCO dataset using UniIR's CLIP-SF and BLIP-FF as retrievers. Column $k \in \{0, 1, 5\}$ shows the number of few-shot examples in each run. For Inception Score (IS), its Standard Deviation (SD) is in parenthesis.

	k	FID ↓	CLIP Score ↑	IS ↑ (SD)
(1a) CLIP-SF	1	13.53	26.72	4.48 (0.16)
(1b) BLIP-FF	1	13.06	27.07	4.28 (0.21)
(2) LaVIT	0	117.81	21.83	**7.74 (0.58)**
(2a) CLIP-SF + LaVIT	1	36.80	28.00	4.74 (0.24)
	5	**30.16**	27.55	4.44 (0.21)
(2b) BLIP-FF + LaVIT	1	36.68	**28.22**	4.31 (0.16)
	5	30.24	27.50	4.36 (0.36)
(2c) Random + LaVIT	1	44.88	23.72	4.39 (0.27)
	5	38.89	26.15	4.13 (0.18)

Table 6: Image generation with LaVIT on the Fashion200k's dataset using UniIR's CLIP-SF and BLIP-FF as retrievers. Column $k \in \{0, 1, 5\}$ represents the number of few-shot examples for each experiment. For Inception Score (IS), its Standard Deviation (SD) is reported in parenthesis.

Table 5 shows the image generation results for the LaVIT and Emu2-Gen (denoted as Emu-G in the table) generator models on the MSCOCO dataset. As expected, both LaVIT with a 7b backbone LLM and Emu2-Gen with a 33b backbone LLM learn from in-context examples and generate better images when text-image pair examples are included in the prompt. However, the amount of improvement is different for each model and it varies across different few-shot settings. While Emu2-G is more effective zero-shot, LaVIT's superior in-context learning ability significantly improves its few-shot generation and makes it the more effective model in all few-shot settings.

Both models generate their best results with a single most relevant RAG few-shot example ($k = 1$), regardless of the retriever model. Further increasing the number of RAG examples to $k = 5$ has a negligible impact, modestly degrading the image fidelity in favor of slightly higher CLIP and/or Inception scores (IS). However, the two models behave differently for random few-shot examples; while LaVIT is still mostly effective with randomly selected examples, Emu2-G seems to get confused by too many random examples and its $k = 5$ random results are much worse than its zero-shot baseline.

Table 6 presents the image generation results for LaVIT on the Fashion200k dataset. As expected, FID and CLIP Score show trends similar to those in image captioning, with random and RAG examples incrementally enhancing the fidelity of generated images to the ground truth (lower FID) and improving the alignment of generated images with caption queries (higher CLIP Score). However, the Inception Score (IS) is highest in the zero-shot setting. This is due to the low diversity of the Fashion200k dataset; generating more faithful images in this context leads to lower diversity among the generated images, resulting in a lower IS. Another notable observation is that, with RAG examples, LaVIT achieves a CLIP Score higher than those of the retrievers. This suggests that the generated images align better with the caption queries than the images retrieved from the dataset. Please see Appendix E for sample visualizations.

5.2.1 Effect of Sampling

For each image in M-BEIR's MSCOCO test set, we have sampled a single caption (Set 1 Captions in Table 7). To study the sensitivity to sampling, we repeated the sampling process one more time (Set 2 Captions in Table 7). The two sample sets have only 1030 captions in common but both of them cover all 5k distinct images in the dataset. We used CLIP-SF + LaVIT in this ablation study as the retriever-generator pipeline.

As shown in Table 7, FID and CLIP Score for all k values are very similar between the two runs with less than one percent difference. However, the reported Inception Scores and their standard deviations vary slightly more, but still within the confidence interval of one another. We believe this difference is related to the inherent uncertainty of the metric when applied to small dataset sizes.

	k	FID ↓	CLIP Score ↑	IS ↑ (SD)
Caption Set 1	0	64.80	26.27	16.38 (1.27)
	1	23.39	30.42	24.11 (1.09)
	5	24.14	31.09	24.52 (1.08)
Caption Set 2	0	64.46	26.08	15.83 (0.86)
	1	23.24	30.38	24.78 (1.97)
	5	24.22	31.03	24.58 (0.75)

Table 7: Image generation with CLIP-SF + LaVIT on two sets of 5k captions sampled from the MSCOCO dataset. Column $k \in \{0, 1, 5\}$ represents the number of few-shot examples for each experiment.

6 Conclusion

We introduced UniRAG as a model agnostic retrieval augmentation technique for Multi-Modal (MM) tasks with image and text modalities and evaluated it on image-to-text and text-to-image tasks. We utilized the LLaVA, Gemini-Pro, and GPT-4o models to generate captions for input images in zero-shot and few-shot (with RAG) settings. Similarly, we employed the LaVIT and Emu2-Gen models to generate images from input captions. In the RAG few-shot setting, we leveraged UniIR's CLIP-SF and BLIP-FF models to retrieve relevant image-text pairs and included them as in-context examples. To further showcase the effectiveness of RAG few-shot examples, we also compared them against randomly selected few-shot examples.

Our experimental results on the MSCOCO dataset from M-BEIR indicated that all models, regardless of their zero-shot baseline effectiveness, improved after being exposed to in-context examples. However, the best performance was achieved only when relevant retrieved examples were included in the prompts, rather than random text-image pairs. In fact, for the LLaVA, GPT-4o, and Emu2-Gen models, adding five random examples decreased generation quality below their zero-shot baselines. In contrast, in the RAG few-shot setting, increasing the number of examples from one to five either improved generation quality (for Gemini-Pro and GPT-4o) or had no effect (for LaVIT and Emu2-Gen). For LLaVA, while adding more RAG in-context examples reduced its effectiveness, it still performed significantly better than both the zero-shot and random few-shot prompts. Our experiments on five MM-LLMs confirm that UniRAG with UniIR retrievers is a model-agnostic effective way to improve the generation quality of pretrained MM-LLMs by retrieving relevant few-shot examples at inference time.

To assess UniRAG's effectiveness with uncommon domain-specific entities, we evaluated it on the Fashion200k dataset. Our results showed that UniRAG is essential to improve generation quality in scenarios where the generator model has limited implicit knowledge about the entities.

7 Limitations

While we open-sourced UniRAG, its usage is still bound by the licenses and the usage agreements of proprietary GPT-4o and Gemini-Pro models used for inference. Furthermore, one of our open-source models LaVIT, is licensed under the LaVIT Community License which requires specific license requests for commercial use.

This work uses English-only MSCOCO and Fashion200k datasets for evaluation and experiments with retriever and generator models that are primarily trained on English corpus. Thus, the effectiveness of our proposed method for non-English low-resource languages remains unexplored. Furthermore, relevance is the only criterion used while retrieving candidate pairs for in-context examples. Deployment of this method in real-life applications requires candidate rankings based on facts, potential harms, biases and other important factors for responsible AI, rather than solely relying on relevance.

Additionally, generating images of real people would require extra caution and extensive auditing since it has significant privacy and security implications. Finally, further research is required to investigate the generalizability of the retriever-guided generation to out-of-domain retrieval. For instance, employing an entity-centric evaluation dataset, which the retriever has not been trained on, could more effectively demonstrate the advantage of retrieval augmentation for multi-modal inference.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. *Preprint*, arXiv:1607.08822.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. 2022. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. *Preprint*, arXiv:2112.04426.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022a. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. 2022b. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*.

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

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. 2023. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*.

Mohammad Dehghan, Mohammad Alomrani, Sunyam Bagga, David Alfonso-Hermelo, Khalil Bibi, Abbas Ghaddar, Yingxue Zhang, Xiaoguang Li, Jianye Hao, Qun Liu, Jimmy Lin, Boxing Chen, Prasanna Parthasarathi, Mahdi Biparva, and Mehdi Rezagholizadeh. 2024. EWEK-QA : Enhanced web and efficient knowledge graph retrieval for citation-based question answering systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14169–14187, Bangkok, Thailand. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *Preprint*, arXiv:2002.08909.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. 2023. 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.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. 2024. Visual hallucinations of multi-modal large language models. *arXiv preprint arXiv:2402.14683*.

Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, et al. 2023. Unified language-vision pretraining with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Jian Li and Weiheng Lu. 2024. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Preprint*, arXiv:2201.12086.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2024. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11238–11254, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bingshuai Liu, Chenyang Lyu, Zijun Min, Zhanyu Wang, Jinsong Su, and Longyue Wang. 2023a. Retrieval-augmented multi-modal chain-of-thoughts reasoning for large language models. *arXiv preprint arXiv:2312.01714*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. *Preprint*, arXiv:2102.12092.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. 2023. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, et al. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. *Preprint*, arXiv:1411.5726.

Keheng Wang, Feiyu Duan, Sirui Wang, Peiguang Li, Yunsen Xian, Chuantao Yin, Wenge Rong, and Zhang Xiong. 2023a. Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering. *arXiv preprint arXiv:2308.13259*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2023. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2022. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling autoregressive models for content-rich text-to-image generation. *Preprint*, arXiv:2206.10789.

Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. 2023. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Appendix A Prompt Templates

This section shows the generic zero-shot and model-specific few-shot prompt templates for the caption generation task.

```
You are an intelligent helpful assistant AI that is an expert in
generating captions for provided images.
I have provided you an image. Generate the caption for this image
based on your understanding of the image.
Only respond with the caption; do not say any other words or explain.
```

Figure 2: Zero-shot prompting for caption generation.

```
{{merged images}}

You are an intelligent helpful assistant AI that is an expert in
generating captions for provided images.

I have provided you {image_num} images. The following {caption_num}
captions correspond to the first {caption_num} images.

{{captions}}

Generate the caption for the last remaining image based on your
understanding of the last image and provided image-caption examples.
Only respond with the caption; do not say any other words or explain.
```

Figure 3: Few-shot prompting for caption generation with LlaVA.

```
I have provided you {num} image(s) and their corresponding caption(s)
as example(s) and one last image without a caption.

Generate the caption for the last remaining image based on your
understanding of the last image and provided image-caption examples.
Only respond with the caption string; do not say any other words or
explain.

{{image-caption-pairs}}
```

Figure 4: Few-shot prompting for caption generation with Gemini-Pro.

```
You are an intelligent helpful assistant AI that is an expert in
generating captions for provided images.

{{image-caption-pairs}}

I have provided you {num} image(s) and their corresponding caption(s)
as example(s) and one last image without a caption.

Generate the caption for the last remaining image based on your
understanding of the last image and provided image-caption examples.
Only respond with the caption; do not say any other words or explain.
```

Figure 5: Few-shot prompting for caption generation with GPT-4o.

Appendix B Proprietary Models Cost

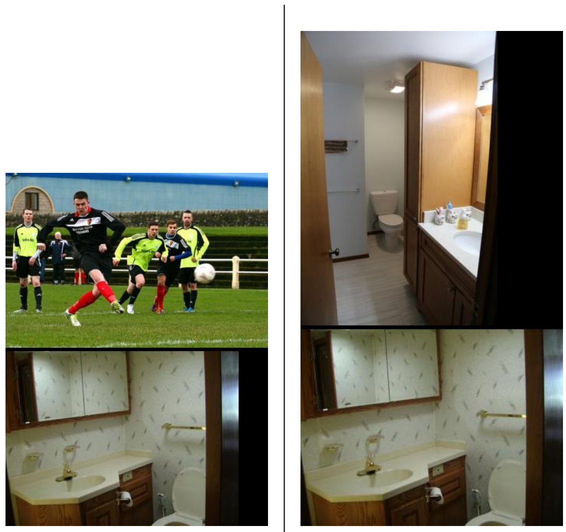
We used OpenAI and Vertex APIs for generating captions with GPT-4o and Gemini-Pro models, respectively. Table 8 illustrates the API pricing information for both models. Overall, for the evaluation of the MSCOCO test set which includes 5k image-text pairs, our average cost estimates per run are approximately 45 USD for the GPT-4o model and 35 USD for the Gemini-Pro model. Image captioning experiments in this paper cost about 560 USD in total (seven runs for each model).

Model	Input		Output
	Text	Image	Text
GPT-4o	\$0.005	\$0.015	\$0.005
Gemini-Pro	\$0.000125	\$0.0025*	\$0.000375

Table 8: Model API pricing for per 1000 tokens. Above mentioned rates are in terms of USD. $*$ - Gemini API call for image input presents cost per image.

Appendix C Generation Visualization

This section shows a sample merged image for few-shot caption generation with LLaVA.



(a) Random (b) CLIP-SF

Figure 6: A sample image merging for LLaVA prompts. Few-shot images are vertically merged into the query-image for (a) random and (b) CLIP-SF retrievers with $k = 1$ examples.

Appendix D Reported Metrics

In our evaluations, we report commonly used metrics for both tasks. For caption generation, we use BLEU (1-4) (Papineni et al., 2002) and ROUGE (Lin, 2004), which measure the precision and recall of n-grams in the generated captions compared to ground-truth captions. We also employ CIDEr (Vedantam et al., 2015), which assesses cosine similarity between n-gram vectors from the generated and ground-truth captions, and SPICE (Anderson et al., 2016), which compares the semantic content of the generated and ground-truth captions using scene-graph tuples.

For image generation, we use FID (Heusel et al., 2017) to measure the KL divergence between the feature vector distributions of generated and ground-truth images; a lower FID indicates greater similarity. Additionally, we calculate the CLIP Score (Hessel et al., 2021), which measures cosine similarity between CLIP's visual and text embeddings for a generated image and its caption query. Lastly, we include the Inception Score (IS) (Salimans et al., 2016) and its Standard Deviation (SD) as another quality indicator for generated images. A higher IS indicates more diversity and confident classification of the generated images. However, IS provides a weaker signal compared to the other two metrics for image generation, as it is less accurate for small datasets.

Appendix E Fashion200k Dataset Visualization

Tables 9 and 10 visualize sample caption and image generation for the Fashion200k dataset.

	Image Query Zeroshot	Top retrieved image-caption pair ($k = 1$) CLIP-SF	BLIP-FF	Random
Prompt				
		Multicolor fine sandy blazer red.	Red petite refined one button blazer.	Dp curve black collared shirt dress.
LLaVA	Effortless style in a vibrant red blazer.	Red blazer with a white collar.	Red blazer with a high neckline.	Red blazer with black collar.

Table 9: Sample caption generation with LLaVA model on the Fashion200k dataset in zero-shot and one-shot settings. The "Prompt" row shows the zero-shot image query as well as retrieved image-caption pairs from CLIP-SF, BLIP-FF and random selection that are included in the prompt as in-context examples.

	Caption Query Zeroshot	Top retrieved image-caption pair ($k = 1$) CLIP-SF	BLIP-FF	Random
Prompt	-			
	Black asymmetric overlay dress.	Black sheer panel dress.	Black asymmetric panel dress.	Blue joanne cropped tailored trousers.
LaVIT				

Table 10: Sample image generation with LaVIT on Fashion200k dataset in zero-shot, and one-shot settings. The "Prompt" row shows the zero-shot caption query as well as retrieved image-caption pairs from CLIP-SF, BLIP-FF and random selection that are included in the prompt as in-context examples.