Unified Embeddings for Multimodal Retrieval via Frozen LLMs

Ziyang Wang[1*] **Heba Elfardy**[2] **Markus Dreyer**[2] **Kevin Small**[2] **Mohit Bansal**[1,2]
[1]UNC Chapel Hill [2]Amazon
{ziyangw,mbansal}@cs.unc.edu
{helfardy,mddreyer,smakevin,mobansal}@amazon.com

Abstract

In this work, We present **Uni**fied Embeddings for **Mu**ltimodal **R**etrieval (UNIMUR), a simple but effective approach that embeds multimodal inputs and retrieves visual and textual outputs via frozen Large Language Models (LLMs). Specifically, UNIMUR jointly retrieves multimodal outputs via unified multimodal embedding and applies dual alignment training to account for both visual and textual semantics. Thus, unlike previous approaches, UNIMUR significantly reduces LLM's modality bias towards generating text-only outputs. Meanwhile, the proposed unified multimodal embedding mitigates the inconsistency between visual and textual outputs and provides coherent multimodal outputs. Empirically, UNIMUR also achieves strong image/text retrieval ability outperforming existing approaches on zero-shot multimodal response retrieval on MMDialog, improving the overall R@1 by 6.5% while boosting the image retrieval rate and having better cross-modal consistency on multimodal outputs. UNIMUR also achieves 2.4% and 3.9% improvement on context-based image retrieval tasks on MMDialog and VisDial respectively when compared to previous approaches, validating its generalization ability across multiple tasks.

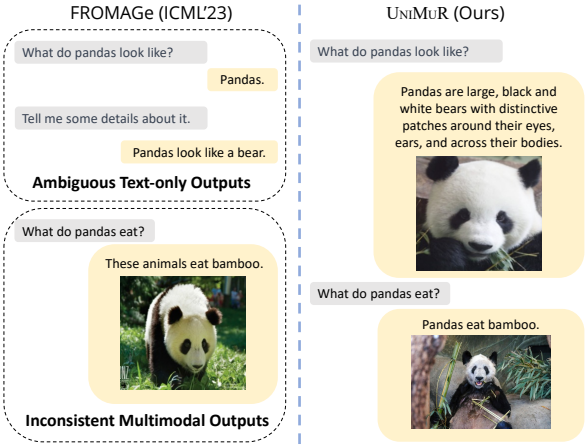

Figure 1: Comparison between FROMAGE baseline and our proposed UNIMUR method. As shown in the top of the figure, UNIMUR is able to more frequently retrieve visual outputs compared to FROMAGE which has a stronger bias to produce text-only outputs. Thus, we leverage unified multimodal embeddings to reduce the ambiguity of text-only outputs with the help of multimodal information. UNIMUR also retrieves more informative textual outputs which align with the visual outputs. Additionally, as shown at the bottom of the figure, via joint retrieval of visual and textual outputs, UNIMUR reduces the inconsistency in multimodal outputs (UNIMUR retrieves the image of a "panda eating bamboo" while the baseline model retrieves a non-specific picture of panda).

1 Introduction

Trained on massive text corpora sourced from the Internet, large language models (LLMs) have showcased remarkable capabilities, ranging from generating human-like dialogue to answering complex queries posed by users (Rae et al., 2021; Touvron et al., 2023; Chowdhery et al., 2022; Zhang et al., 2022; ChatGPT, 2022). One limitation of most widely available state-of-the-art LLMs–with some exceptions–is that they focus on text-only interactions and do not utilize visual information. However, beyond language, visual information is a fundamental signal through which humans perceive and engage with their surroundings. Consequently, building LLMs that can embed and retrieve visual and textual information is crucial for enhancing the user experience when interacting with the model.

One approach for enabling multimodal inputs and outputs with LLMs is to train a Multimodal LLM (MLLM) with large-scale multimodal data (Alayrac et al., 2022; Yu et al., 2022; Gao et al., 2023; Zhu et al., 2023; GPT-4, 2023). However, this approach requires costly large-scale pretraining and primarily focuses on learning multimodal input embeddings relative to optimizing for efficient retrieval or generation of multimodal outputs.

*Work conducted during an internship at Amazon.

Recent work propose to instruct the frozen LLM to generate a special retrieval token to retrieve or generate an image given multimodal inputs (Koh et al., 2023a,b). Despite their success, due to the LLM's extensive pretraining on text-only data, this approach generally exhibits a strong bias towards generating text tokens and not the special image token, resulting in a low prevalence of responses with visual information. As shown in Figure 1, given the question "What do pandas look like", such approaches frequently give a text-only answer "A panda" instead of also showing an image of a panda to illustrate what it looks like.

In this work, we propose **Uni**fied Embeddings for **Mu**ltimodal **R**etrieval (UNIMUR), which aims to mitigate this modality bias by efficiently retrieving multimodal outputs via a unified embedding which aligns to both visual and textual semantics. UNIMUR utilizes a simple yet effective approach for embedding multimodal inputs and retrieving multimodal outputs via frozen language models. Unlike previous methods, UNIMUR maps the LLM output embeddings to the unified multimodal embeddings for retrieving both visual and textual outputs. To train the unified multimodal embedding, we propose a dual alignment training strategy that matches the unified multimodal embedding to both visual and textual semantics.

UNIMUR has three primary strengths: (1) It significantly reduces the text-only bias resulting in more frequent retrieval of multimodal outputs and enrich the text-only outputs with visual information. As shown in Figure 1, given the question *"what do pandas look like"*, UNIMUR is able to retrieve a more informative than the baseline multimodal response that contains both visual and textual descriptions. Experimental results show that UNIMUR significantly increases the number of dialogue turns that also include retrieved visual responses. (2) UNIMUR retrieves multimodal outputs with better cross-modal consistency via its joint retrieval pipeline. As shown in Figure 1, given the question *"what do pandas eat"*, UNIMUR is able to retrieve the textual response *"Pandas eat bamboo."* together with an image that matches the text (instead of retrieving a non-specific image with pandas). Quantitative results show that UNIMUR achieves higher CLIP-similarity a FROMAGE baseline by 2.6% between its visual and textual outputs. (3) We empirically show that our dual-alignment training strategy for the unified multimodal embedding improves the retrieval for both image and text candidates, which indicates that the knowledge sharing between visual and textual information is useful for retrieval performance on both ends.

To validate the effectiveness of UNIMUR, we first evaluate its performance on the zero-shot multimodal response retrieval task using the MMDialog dataset (Feng et al., 2023). Secondly, we evaluate performance on the contextual image retrieval and dialogue-to-image retrieval tasks on both multimodal chitchat and image-centric dialogue datasets. On MMDialog, experimental results show that UNIMUR significantly reduces the text-only output bias with stronger retrieval performance in the zero-shot setting. UNIMUR also achieves better results on the contextual image retrieval and dialogue-to-image retrieval tasks, indicating its improvements generalizing to multiple tasks.

To summarize, our contributions are:

- We propose UNIMUR, a simple but effective approach that embeds multimodal inputs and retrieves multimodal outputs via frozen language models;
- We apply a dual-alignment training strategy to jointly retrieve the visual and textual outputs via a unified multimodal embedding that significantly reduces the text-only response bias and retrieves multimodal outputs with increased cross-modal consistency;
- We empirically show that UNIMUR achieves better performance on a zero-shot multimodal response retrieval task as well as better results on multiple zero-shot image retrieval tasks.

2 Related Work

Large Language Models. There have been significant recent advancements in the field of large language models (LLMs). Models with parameter counts exceeding 100B, such as GPT-3 (Brown et al., 2020) have demonstrated remarkable proficiency across a wide range of tasks and gained popularity well beyond the research community. Subsequently, a number of follow-up works have been introduced, aiming to enhance different aspects of LLMs' capabilities (e.g., scaling the model size and pretraining data, and improving fine-tuning objectives) (Rae et al., 2021; Touvron et al., 2023; Thoppilan et al., 2022; Chowdhery et al., 2022; Zhang et al., 2022; ChatGPT, 2022). These LLMs primarily aim to tackle different tasks

in zero- and few-shot manner. In this work, we leverage the zero-shot generalization ability of the pretrained LLMs to tackle multiple diverse downstream multimodal tasks.

Embedding Multimodal Inputs Using LLMs. For LLMs to understand visual input, previous works propose to train a mapping function (module) to convert the visual representation into text space that can be directly processed by LLMs (Li et al., 2022; Tsimpoukelli et al., 2021; Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023a; Huang et al., 2023a; Lv et al., 2023; Eichenberg et al., 2021; Yu et al., 2023; Berrios et al., 2023; Aghajanyan et al., 2022; Yi-Lin Sung, 2022; Wang et al., 2022; Cho et al., 2021; Ilharco et al., 2020; Wu et al., 2023; Huang et al., 2023b; Zhang et al., 2023). Specifically, *Frozen* (Tsimpoukelli et al., 2021) trains a vision encoder to represent each image as a sequence of continuous embeddings as input to LLMs. LIMBeR (Merullo et al., 2022) shows that the image representations from vision models can be transferred as continuous prompts to frozen LMs by training only a single linear projection. BLIP-2 (Li et al., 2023) utilizes Q-Former to align the visual features with an LLM while LLaVA (Liu et al., 2023a) injects visual features into the language model by treating image tokens as a foreign language, and using conversations generated by GPT-4 for fine-tuning. As opposed to these methods, our proposed UNIMUR method focuses on jointly embedding multimodal inputs and retrieving multimodal outputs with minimal training cost.

Producing Multimodal Outputs Using LLMs. Recently, several works have also explored the potential of producing multimodal outputs via LLMs (Sun et al., 2023; Koh et al., 2023b,a; Yasunaga et al., 2023; Liu et al., 2023b). Specifically, FROMAGE (Koh et al., 2023b) trains a multimodal language model capable of generating free-form text interleaved with retrieved images. GILL (Koh et al., 2023a) extends the FROMAGE model with image generation ability. While these models successfully produce multimodal outputs with frozen LLMs, they have two main limitations: (1) due to the LLM's extensive pretraining on the textual corpus, these models suffer from text-only bias while generating output responses, and (2) the textual and visual output is produced by separate processes, which can incur inconsistencies between the multimodal outputs. UNIMUR mitigates these limitations by utilizing a unified multimodal embedding to jointly retrieve visual and textual outputs.

3 Methodology

In this section, we present UNIMUR, a general approach based on frozen LLMs and image-text pretrained models. The training pipeline of our proposed approach is illustrated in Figure 2. We propose two alternating steps to embed multimodal inputs and retrieve multimodal outputs via frzoen LLMs. As shown in the left part of Figure 2, in the image-to-text training step, we train a linear mapping layer that maps the image into LLM's input space in order to access the multimodal understanding ability of the LLM. In the dual alignment training step, we propose to match the visual and textual semantics with the unified multimodal embedding, shown in the right part of Figure 2. During inference, UNIMUR jointly retrieves multimodal outputs via the trained unified multimodal embedding. Below, we discuss the different components of our UNIMUR approach in more detail.

3.1 Pretrained Models

Large Language Models (LLMs): To leverage the knowledge from large-scale language pretraining, UNIMUR utilizes an auto-regressive LLM p_θ and keeps the LLM's parameters θ frozen. Given the input text T, the LLM first extracts a sequence of input tokens $(t_1, \dots t_M)$ via its tokenizer. These LLMs are trained to maximize the log likelihood of the input token sequence by conditioning the next token t_m on all previous tokens $(t_1, \dots, t_{m-1})$. LLMs are considered as strong tools for embedding complex input context with the potential to generate useful embeddings for multimodal output retrieval. Specifically, we leverage the last output embeddings H_θ of the LLM as the generated embeddings for further multimodal output training.

Image-Text Pretrained Models: To represent the visual and textual semantics, we leverage the image-text pretrained model CLIP (Radford et al., 2021), which is a dual-stream image-text model that was pretrained with a contrastive loss on 400 million image-text pairs. It utilizes a GPT-style (Radford et al., 2019) Transformer-based text encoder and a VisionTransformer (ViT) image encoder (Dosovitskiy et al., 2021). Specifically, given an image i and text t, we extract the visual $v_\phi(i) \in \mathbb{R}^c$ and textual $s_\phi(t) \in \mathbb{R}^c$ semantic repre-

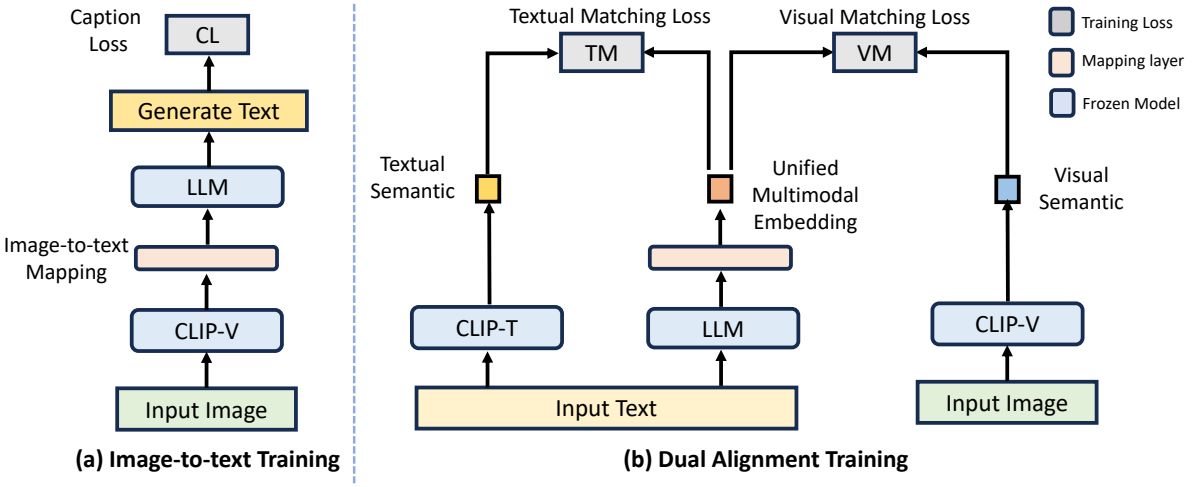

Figure 2: UNIMUR is trained in two alternating steps: (a) The image-to-text training step learns an image-to-text mapping layer via image caption objective, enabling multimodal input; and (b) the dual alignment training maps the LLM output embedding to a unified multimodal space. The unified multimodal embedding is trained to perform visual and textual matching by aligning the visual and textual semantics extracted by the CLIP encoders.

sentations.

3.2 Image-to-Text Training

To embed multimodal inputs via LLMs, we aim to map the image into the LLM input space (i.e. text space). Specifically, we first use CLIP visual encoder to extract the visual embedding $\hat{v}_\phi(i) \in \mathbb{R}^c$ of the given image i. Then, following Merullo et al. (2022) and Koh et al. (2023b), we learn a linear mapping $\mathbf{W}_{i2t} \in \mathbb{R}^{c \times d}$ from the image's visual embeddings $\hat{v}_\phi(i)$ into the LLM's input space as $\hat{v}_\phi(i)^T \mathbf{W}_{i2t} \in \mathbb{R}^d$. This allows the model to translate the visual inputs to "language-like" tokens that can directly be processed by the LLM. As shown in Figure 2(a), to train this mapping layer, we apply the image captioning objective which generates text tokens within the textual caption conditioned on the visual prefix. The visual prefix (i.e., "language-like" tokens) is the output of the image-to-text mapping layer, which is prepended to the textual caption. The log-likelihood of textual caption t conditioned on its image i is:

$$l_c(i,t) = \sum_{m=1}^{M} \log p_\theta \left(t_m \mid \hat{v}_\phi(i)^T \mathbf{W}_{i2t}, t_1, \ldots, t_{m-1}\right) \quad (1)$$

Then, the image captioning loss $\mathcal{L}_{cap}$ is the negative log-likelihood of all samples in a batch of N image-text pairs:

$$\mathcal{L}_{cap} = -\frac{1}{N} \sum_{i=1}^{N} l_c\left(i_j, t_j\right) \quad (2)$$

By applying this image-to-text mapping, we convert a set of multimodal inputs to "text-only" inputs and feed it into the LLM, enabling the LLM to embed complex multimodal inputs. Since our training for multimodal inputs and outputs is modularized, our image-to-text mapping is model-agnostic, providing the flexibility to incorporate any advanced mapping strategies and achieve better performance in the future.

3.3 Dual Alignment Training

Next, we describe how we train UNIMUR to retrieve multimodal outputs consisting of paired image-text data. In order to avoid the text-only output bias of previous methods (Koh et al., 2023a,b), which used separate processes for visual and textual retrieval, we optimize a unified embedding to jointly retrieve visual and textual outputs. Specifically, we map the LLM's last output embedding $H_\theta \in \mathbb{R}^p$ to a unified multimodal space with a linear mapping layer $\mathbf{W}_{t2m} \in \mathbb{R}^{p \times q}$ and obtain the unified multimodal embedding $e = H_\theta^T \mathbf{W}_{t2m} \in \mathbb{R}^q$. By applying the unified multimodal embedding, we improve the cross-modal consistency of the multimodal outputs and mitigate the potential inconsistency caused by the separate image and text retrieval processes.

As shown in Figure 2(b), to further alleviate the modality bias of the LLM output, we adopt a dual alignment training (DAT) method that aligns the unified multimodal embedding with both visual and textual semantics. Specifically, we utilize two training objectives: visual matching (VM) loss and textual matching (TM) loss. For visual matching

loss, we aim to align our unified multimodal embeddings with the visual semantics provided by CLIP visual encoder for image retrieval ability, shown in the right part of Figure 2(b). Thus, we apply a contrastive learning objective with the InfoNCE loss (Oord et al., 2018), a type of contrastive loss function which is widely used for representation learning. Note that the dimensionality of unified multimodal embeddings is equivalent to visual/textual semantics hence we are able to directly apply matching objectives without additional mappings. Given the input text caption t and image i, we calculate the normalized cosine similarity for the visual semantics $v_\phi(i)$ and the unified multimodal embeddings for the input text e_t as:

$$\text{sim}(e_t, i) = \frac{e_t v_\phi(i)^T}{\|e_t\| \, \|v_\phi(i)^T\|}. \tag{3}$$

We minimize the InfoNCE loss in a symmetric manner over a batch of N text-image pairs and contrast over the unified multimodal embedding for the text caption and the visual semantic of the image (e_j, v_k) (here e stands for e_t, v stands for $v_\phi(i)$), where each paired example is considered as a positive pair, and other in-batch examples as negatives:

$$\begin{aligned} \mathcal{L}_{\text{m2v}} &= -\frac{1}{N}\sum_{j=1}^{N}\left(\log \frac{\exp\left(\text{sim}\left(e_j, v_k\right)/\tau\right)}{\sum_{k=1}^{N}\exp\left(\text{sim}\left(e_j, v_k\right)/\tau\right)}\right) \\ \mathcal{L}_{\text{v2m}} &= -\frac{1}{N}\sum_{k=1}^{N}\left(\log \frac{\exp\left(\text{sim}\left(v_k, e_j\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}\left(v_k, e_j\right)/\tau\right)}\right) \end{aligned} \tag{4}$$

$$\mathcal{L}_{vm} = \mathcal{L}_{\text{v2m}} + \mathcal{L}_{\text{m2v}} \tag{5}$$

For textual matching loss, as shown in the left part of Figure 2(b), the target is to preserve the language understanding ability of the unified multimodal embedding and prevent the modality bias created by single visual matching objective. To this end, we align the textual semantics with the unified multimodal embedding. Since the domain gap between the LLM output embedding and textual semantics is limited, inspired by VLKD (Dai et al., 2022), we employ a stricter alignment objective between multimodal embedding e and textual semantics $s_\phi(t)$. Specifically, given the textual caption t, we utilize Mean Square Error (MSE) to minimize the $\mathcal{L}_2$ distance between e_t and $s_\phi(t)$:

$$\mathcal{L}_{tm} = \|e_t - s_\phi(t)\|_2^2. \tag{6}$$

In summary, the overall training objective is:

$$\mathcal{L} = \mathcal{L}_{cap} + \lambda_1 \mathcal{L}_{vm} + \lambda_2 \mathcal{L}_{tm}, \tag{7}$$

where λ_1 and λ_2 are hyper-parameters which define the relative weights of the visual and text matching losses.

3.4 Retrieving Multimodal Outputs

During inference, UNIMUR retrieves both visual and textual outputs using the unified multimodal embeddings given the input contexts. Specifically, we first map the image to text space via the image-to-text mapping layer $\mathbf{W}_{i2t}$ and feed the result to the LLM. We then map the LLM's last output embedding to the unified multimodal embedding e via the linear mapping layer $\mathbf{W}_{t2m}$. Given the multimodal candidate pool, we extract the visual embeddings via CLIP encoder. For textual candidates, we directly use the LLM's average input embeddings of textual candidates as the candidate embeddings. We then concatenate the candidate visual and textual embeddings to a candidate pool and utilize the unified multimodal embedding e to retrieve the most relevant multimodal candidates from the pool. Specifically, we leverage cosine similarity to calculate the relevance between the unified multimodal embedding and multimodal candidates. We then select the most relevant candidates with the highest similarities.

4 Experiments

4.1 Tasks and Evaluation Metrics

Multimodal Response Retrieval. We first evaluate our model on a multimodal response retrieval task which requires the model to embed the multimodal dialogue context and retrieve the visual and textual responses for current dialogue turn. For this, we test the zero-shot performance on MMDialog (Feng et al., 2023), a large-scale multi-turn dialogue dataset containing multi-modal open-domain conversations derived from human-human chat content in social media. For each turn, we retrieve the top-2 samples from the multimodal candidate pool. Since the conversation turns in MMDialog are in two categories – text-only and visual+text responses – we first retrieve the top-1 text candidate as the textual utterance. Then, we retrieve the most relevant candidate from the remaining candidate pool and output the image responses.[1] Thus, UNIMUR is capable of retrieving image responses to facilitate the text-only dialogue without an additional intent prediction module (Feng et al., 2023).

[1] If the top-2 samples are both textual candidates, we output the top-1 candidate as the textual utterance.

We evaluate the multimodal response retrieval performance based on three aspects: (1) the extent of text-only bias within the outputs; (2) the accuracy of the retrieved outputs; and (3) the consistency of the multimodal text and image outputs. First, to quantify the model's text-only bias, we report the rate at which the model retrieves image candidates when the ground truth response contains visual responses (**image retrieval rate**). Then, to show the correctness of the retrieved outputs, we report the standard recall rate for both visual and textual response retrieval using **R@1**.[2] We also report the **overall R@1** on all responses to show the general multimodal retrieval performance. To test the semantic consistency of visual and textual outputs, we report the average cosine similarity between the CLIP embeddings of the retrieved visual and textual outputs (**CLIP-Sim**).

Contextual Image Retrieval. To evaluate the model's image retrieval ability given a complex multimodal context, we test our model on the contextual image retrieval task. We test the zero-shot image retrieval performance on the dialogue turns that contain visual responses from the MMDialog dataset (Feng et al., 2023). Specifically, given the multimodal dialogue context, we require the model to retrieve the correct image from the visual candidate pool. Importantly, this task can be considered as the image retrieval part of the previous multimodal response retrieval task while given the ground truth modality information of the dialogue turns. Thus, the performance of contextual image retrieval is capable of showing the model's image retrieval ability regardless of the modality bias. We leverage the standard recall rates $R@1$,$R@5$, and $R@10$ as evaluation metrics.

Dialogue-to-Image Retrieval. To further evaluate UNIMUR on different types of dialogue data, we test our model on the image-centric dataset - Visual Dialog (VisDial (Das et al., 2017)). We report the zero-shot performance on the dialogue-to-image retrieval task, which requires the model to retrieve the correct image given a conversation about it. This task tests the model's ability to embed complex contexts and retrieve the most relevant image given the dialogue context. Here we again use the standard recall rates $R@1$,$R@5$, and $R@10$ as evaluation metrics.

[2]We only consider the first visual response in each turn as ground truth.

4.2 Training Data and Implementation Details

Following (Merullo et al., 2022; Koh et al., 2023b), we train UNIMUR on the Conceptual Captions (CC3M) dataset (Sharma et al., 2018) consisting of 3.3 million image-text pairs. To improve the retrieval abilities of auto-regressive LLM, we add a special [RET] token at the end of each input context to represent embeddings for multimodal retrieval (Koh et al., 2023b).

We utilize the publicly available OPT model (Zhang et al., 2022) with 6.7B parameters as our LLM. Past work mentions that findings at the 6.7B scale are large enough to exhibit the zero-shot learning abilities that we are interested in (Koh et al., 2023b; Radford et al., 2019). For the image-text pretraining model, we utilize the pretrained CLIP ViT-L/14 model (Radford et al., 2021) for its ability to produce strong visual/textual semantic information (Wang et al., 2023).

We implemented our model on PyTorch (Paszke et al., 2019) and trained mixed-precision with BFloat16 (Abadi et al., 2016). Since most of the model parameters (98.0%) are frozen, our method is computationally efficient and we only optimize the parameters from two linear mapping layers. We use the Adam (Kingma and Ba, 2014) optimizer with a learning rate 0.0002 and warmup of 100 steps. We set the LLM's input dimension $d = 4096$ (inherited from OPT-6.7B) and the dimension of multimodal embedding as 768. Via simple hyperparameter search, we set the weight of visual matching loss as 1 and textual matching as 10. We train our model with 5 epochs and the training time is less than 16 hours on 4 NVIDIA V100 GPUs.

5 Results

In this section, we present the empirical results of our proposed approach – UNIMUR. We evaluate UNIMUR on 3 different multimodal retrieval tasks; multimodal response retrieval (Section 5.1), contextual image retrieval (Section 5.2), and dialogue-to-image retrieval (Section 5.3).

5.1 Multimodal Response Retrieval

We evaluate UNIMUR's performance on zero-shot multimodal response retrieval task and compare its performance to the recent FROMAGE model (Koh et al., 2023b). For a fair comparison, we leverage the same LLM and CLIP checkpoints for both models. Results show that FROMAGE suffers from severe text-only bias with an image retrieval rate

Method	Image Retrieval Rate (%)	Image R@1	Text R@1	Overall R@1	CLIP-Sim
BLIP-2	18.9	16.8	24.5	22.2	0.1755
FROMAGE	28.2	11.0	17.1	15.3	0.1024
FROMAGE-ppl	28.2	11.0	32.5	25.8	0.1618
UNIMUR (Ours)	**68.3**	**23.2**	**36.1**	**32.3**	**0.1873**

Table 1: Zero-shot multimodal response retrieval results on MMDialog dataset. We use FROMAGE-ppl as a baseline which utilizes a highly time-consuming perplexity-based method for text retrieval. Results show that UNIMUR achieves better performance on all metrics while significantly reducing the text-only bias.

Method	R@1	R@5	R@10
FROMAGE	25.4	25.7	26.0
UNIMUR(Ours)	**27.8**	**28.0**	**28.4**

Table 2: Zero-shot contextual image retrieval results on MMDialog dataset.

of only 28.2%, indicating that most of the visual responses fail to be retrieved by the model leading to a low image $R@1$. For text retrieval with FROMAGE, we first apply the same embedding-based retrieval setting with our approach to search the text utterance and get poor text $R@1 = 17.1$ (shown at the top of Table 1). We then apply the perplexity-based method following Koh et al. (2023b), which computes the perplexity of each context and candidate text sequence prior to selecting the text candidate with the lowest perplexity. While improving the text $R@1 = 32.5$ performance (shown in the middle of Table 1), this perplexity-based text retrieval pipeline is extremely time-consuming ($20\times$ compared to UNIMUR), which makes it sub-optimal for real-world applications.

We also compare our method with recent multimodal LLM (BLIP-2 (Li et al., 2023)). Results show that our proposed approach has much less text-only bias (68.3% Image Retrieval Rate compared to 18.9% of BLIP-2), and also has a significant improvement on both image and text retrieval given complex input context. Furthermore, compared to BLIP-2 (16 A100 GPU * 9 days), our UNIMUR model requires much fewer computational resources (4 V100 GPU * 16 hours), proving its efficiency. We argue that multimodal LLMs like BLIP-2 mainly focus on embedding paired multimodal input and producing text-only outputs, which makes it sub-optimal for processing interleaved multimodal input and retrieving visual and textual outputs.

In contrast, as shown at the bottom of Table 1, UNIMUR achieves better performance on all metrics compared to the existing methods. Of particular note, UNIMUR obtains a 68.2% image retrieve rate, outperforming the FROMAGE approach by 40.1%. This indicates our approach significantly reduces the text-only bias within the LLM output. With a better image retrieve rate, UNIMUR also achieves better image $R@1$, outperforming the FROMAGE model by 12.2%. Note that we show in Section 5.2, given the output modality information (image retrieve rate as 1), UNIMUR still outperforms the baseline model by a significant margin. Meanwhile, compared to the perplexity-based FROMAGE model, we achieve 3.6% improvement on text $R@1$ while using significantly less inference time. Furthermore, UNIMUR shows a 6.5% improvement on overall $R@1$, which indicates that in general, UNIMUR is more powerful in embedding multimodal inputs and retrieving multimodal outputs. UNIMUR also achieves a higher CLIP similarity on its visual and textual outputs in the same dialogue turns, indicating our approach is capable of retrieving visual and textual outputs with better cross-modal consistency.

5.2 Contextual Image Retrieval

To evaluate the image retrieval ability given multimodal input, we also report the zero-shot contextual image retrieval results on the MMDialog dataset in Table 2. Results show that UNIMUR outperforms the baseline FROMAGE approach by 2.4% for R@1, indicating that the unified multimodal embedding is capable of capturing important visual information for image retrieval. This also shows that the UNIMUR's improvement on Image $R@1$ of multimodal response retrieval is not just due to the reduction of modality bias, but also takes advantage of more powerful zero-shot image retrieval ability. One additional observation is that the $R@5$ and $R@10$ of both models are not significantly higher than $R@1$, which may be due to using a zero-shot protocol for these evaluations.

5.3 Dialogue-to-image Retrieval

We evaluate UNIMUR on zero-shot dialogue-to-image retrieval on the Visual Dialog dataset(VisDial). This task requires the model to retrieve the correct image given a complex

Method	R@1	R@5	R@10
CLIP	17.7	38.9	50.2
FROMAGE	20.8	44.9	56.0
UNIMUR (ours)	**24.7**	**49.8**	**60.9**

Table 3: Zero-shot dialogue-to-image retrieval results on VisDial dataset.

	Img Retr. Rate(%)	Img R@1	Text R@1	Overall R@1
VM	**74.3**	22.4	29.8	27.6
TM	44.1	14.2	26.4	22.7
DAT	68.3	**23.2**	**36.1**	**32.3**

Table 4: Comparison of different training strategies; Visual Matching (VM), Textual Matching (TM), and our proposed Dual Alignment Training (DAT). DAT achieves better multimodal response retrieval performance while preserving a rather low modality bias.

dialogue context. As shown in Table 3, UNIMUR outperforms CLIP (Radford et al., 2021) and FROMAGE (Koh et al., 2023b) on all metrics, improving the R@1 by 3.9% compared to FROMAGE baseline. This reveals the generalization ability of UNIMUR given complex text-only dialogue contexts.

6 Analysis

In this section, we further analyze UNIMUR to understand the impact of different model design choices as well as to showcase its capabilities.

6.1 Ablation Study

The Effect of Dual Alignment Training. First, we validate the effectiveness of the dual alignment training strategy in our UNIMUR method. As shown in Table 4, compared to visual matching only (VM) and textual matching only (TM) training, our dual alignment training strategy (DAT) achieves better multimodal output quality while preserving a rather low modality bias. Specifically, although image matching only training obtains a better image retrieve rate, the training is biased to the image domain and has a significant drop in text R@1. Meanwhile, the Image R@1 under dual alignment training is also better than image matching only training, indicating that knowledge

LOSS	Img Retr. Rate(%)	Img R@1	Text R@1	Overall R@1
Info-NCE	55.6	17.7	32.2	28.0
Max-Margin	49.8	15.1	30.7	26.3
MSE (UNIMUR)	**68.3**	**23.2**	**36.1**	**32.3**

Table 5: Comparison of different training objectives for textual matching training. Results show that regression-based objective (MSE) outperforms the contrastive learning objectives.

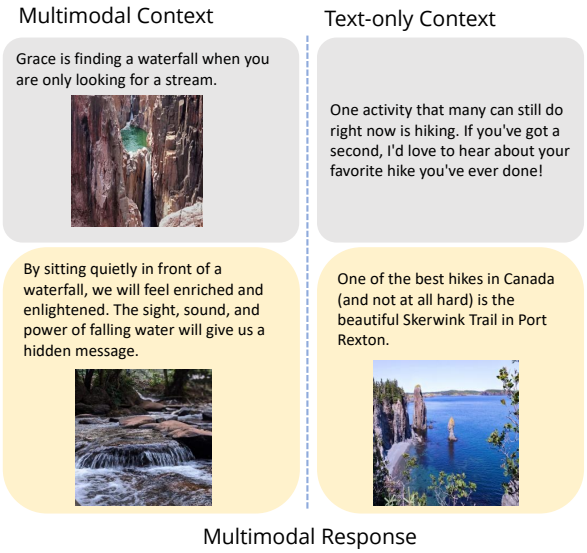

Figure 3: Selected examples from UNIMUR on embedding multimodal input and retrieving multimodal output.

sharing between multimodal data is beneficial for the uni-modal retrieval performance. For textual matching only training, the model suffers from significant text-only bias and has a low image retrieve rate and R@1. Since our target is to jointly retrieve visual and textual outputs, it is crucial to align the unified multimodal embedding to both visual and textual semantics.

Different Training Objective for Text Matching. In Table 5, we compare the different loss choices for our text matching training. Results show that the regression-based objective performs better than the contrastive objective (Info-NCE and Max-Margin). We argue that is because the CLIP text encoder is not powerful enough compared to LLMs and we have to apply a more strict loss function upon text matching training. Meanwhile, another possible reason is that due to the limitation of computational resources, we apply a rather small batch size while training, which is unfavorable for contrastive objectives that highly rely on massive negative samples.

Larger Multimodal Corpora and LM Architectures. As discussed in the implementation detail section 4.2, we follow (Koh et al., 2023b; Merullo et al., 2022) leveraging OPT-6.7B as LLM and Conceptual Caption 3M as training data. Since the LLM is frozen, from a methodological perspective, we can simply scale our approach to larger LM architectures by changing the LLM checkpoints. As shown in Table 6, our approach achieves better VisDial dialogue-to-image retrieval results on

LLM	OPT-1.3B	OPT-7B	OPT-13B
VisDial R@1	16.5	24.7	27.8

Table 6: Comparison of LLMs with different capacity.

larger LLM backbones and indicates its potential to get even better results on LLM over 100B parameters. We also scale up the training dataset using the 12M version of Conceptual Caption. UNIMUR achieves 1.5% performance gain on VisDial dialogue-to-image retrieval using a larger multimodal corpus, proving its generalization ability towards even larger training data.

6.2 Qualitative Analysis

Next, we show some examples of UNIMUR's retrieval results on the MMDialog dataset. As the left side of Figure 3 shows, UNIMUR is capable of embedding multimodal context and retrieving visual and textual responses (in this case, a topic about waterfalls). In addition, UNIMUR is also capable of handling lengthy text-only input and retrieving visual and textual outputs (as shown on the right side of the figure). This last example shows that our model is flexible with different input contexts and is able to retrieve both visual and textual outputs.

7 Conclusion

We present **Uni**fied Embeddings for **Mu**ltimodal **R**etrieval (UNIMUR), a simple yet effective approach that retrieves visual and textual output via unified multimodal embeddings and significantly reduces the text-centric bias from the LLM's output as compared to previous approaches. We empirically show that UNIMUR achieves better zero-shot multimodal response retrieval than state-of-the-art approaches through its joint retrieval process that is capable of retrieving multimodal outputs with better cross-modal consistency. In addition, UNIMUR improves dialogue-to-image retrieval and contextual image retrieval performance to demonstrate its improved performance across multiple tasks.

Limitations

Given multimodal input, we focus on the joint retrieval of visual and textual outputs using frozen large language models. However, given an imperfect candidate pool, retrieval can fail to provide a perfect candidate that matches the input context. We plan to extend our model to multimodal generation. Specifically, given the multimodal input, could we directly generate textual and visual outputs using the unified multimodal embedding? We leave this question for future works.

Ethical Considerations

This paper presents a novel approach for multimodal output retrieval using frozen Large Language Models (LLMs). We leverage LLMs to extract the embeddings for both visual and textual retrieval and not generate any novel visual/textual data. Thus, the proposed method does not introduce additional ethical/social bias given a reliable retrieval candidate pool.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, et al. 2022. Cm3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*.

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.

William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. 2023. Towards language models that can see: Computer vision through the lens of natural language. *arXiv preprint arXiv:2306.16410*.

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

ChatGPT. 2022. [link].

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Enabling multimodal generation on clip via vision-language knowledge distillation. *arXiv preprint arXiv:2203.06386*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh,

Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. 2021. Magma–multimodal augmentation of generative models through adapter-based finetuning. *arXiv preprint arXiv:2112.05253*.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2023. MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7348–7363, Toronto, Canada. Association for Computational Linguistics.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

GPT-4. 2023. https://openai.com/gpt-4.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023a. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*.

Yupan Huang, Zaiqiao Meng, Fangyu Liu, Yixuan Su, Nigel Collier, and Yutong Lu. 2023b. Sparkles: Unlocking chats across multiple images for multimodal instruction-following models. *arXiv preprint arXiv:2308.16463*.

Gabriel Ilharco, Rowan Zellers, Ali Farhadi, and Hannaneh Hajishirzi. 2020. Probing contextual language models for common ground with visual representations. *arXiv preprint arXiv:2005.00619*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. 2023a. Generating images with multimodal language models. *NeurIPS*.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023b. Grounding language models to images for multimodal inputs and outputs.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Yang Yang, Qingyun Li, Jiashuo Yu, et al. 2023b. Internchat: Solving vision-centric tasks by interacting with chatbots beyond language. *arXiv preprint arXiv:2305.05662*.

Tengchao Lv, Yupan Huang, Jingye Chen, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, et al. 2023. Kosmos-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419*.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2022. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing

Liu, Tiejun Huang, and Xinlong Wang. 2023. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

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*.

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

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2023. Unified coarse-to-fine alignment for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2816–2827.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

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

Mohit Bansal Yi-Lin Sung, Jaemin Cho. 2022. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *CVPR*.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2023. Self-chained image-language model for video localization and question answering. *arXiv preprint arXiv:2305.06988*.

Youngjae Yu, Jiwan Chung, Heeseung Yun, Jack Hessel, JaeSung Park, Ximing Lu, Prithviraj Ammanabrolu, Rowan Zellers, Ronan Le Bras, Gunhee Kim, et al. 2022. Multimodal knowledge alignment with reinforcement learning. *arXiv preprint arXiv:2205.12630*.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2023. A simple llm framework for long-range video question-answering.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Models	Frozen	FROMAGE	UNIMUR
VQA Acc	25.5	28.5	29.8

Table 7: VQA results.

Appendix

A Additional Evaluation Tasks

A.1 Visual Question Answering

While the main goal of this paper is to embed interleaved multimodal input and retrieve multimodal outputs, we also show the effectiveness of our approach on visual question answering (VQA) tasks. In Table 7, we show the results of the zero-shot visual question answering task on the VQAv2 dataset (Goyal et al., 2017) (following the FROMAGe (Koh et al., 2023b) setup). Note that our UNIMUR mainly focuses on multimodal retrieval and has no additional training objective related to multimodal reasoning. Still, compared to the baselines (Frozen (Tsimpoukelli et al., 2021) and FROMAGe (Koh et al., 2023b)) that also leverage frozen LLMs, our approach still achieves better VQA results, validating the robustness of the proposed approach.

A.2 Integrating the Unified Embedding with Multimodal Generation Framework

We further extend our method to multimodal generation by simply incorporating the dual alignment training to the recently proposed multimodal generation framework GILL (Koh et al., 2023a) (a contextual image generation method using frozen LLM and Stable Diffusion (Rombach et al., 2021)). We compare our augmented version with the original GILL framework on contextual image generation and contextual image retrieval on the VisDial dataset. Results show that our approach retains strong multimodal generation ability (0.642 vs 0.645 on CLIP-Similarity) while having significant improvement on multimodal retrieval (24.6 vs 21.7 on contextual image retrieval R@1). This indicates our approach is generalizable for different output types including generation with minimal model change.