Improving Context Understanding in Multimodal Large Language Models via Multimodal Composition Learning

Wei Li [1] **Hehe Fan** [1] ✉ **Yongkang Wong** [2] **Yi Yang** [1] **Mohan Kankanhalli** [2]

Abstract

Previous efforts using frozen Large Language Models (LLMs) for visual understanding, via image captioning or image-text retrieval tasks, face challenges when dealing with complex multimodal scenarios. In order to enhance the capabilities of Multimodal Large Language Models (MLLM) in comprehending the context of vision and language, we introduce Multimodal Composition Learning (MCL) for the purpose of mapping or aligning the vision and language input. In particular, we introduce two tasks: Multimodal-Context Captioning (MC-Cap) and Multimodal-Context Retrieval (MC-Ret) to guide a frozen LLM in comprehending the vision and language context. These specialized tasks are crafted to improve the LLM's capacity for efficient processing and utilization of multimodal inputs, thereby enhancing its proficiency in generating more accurate text or visual representations. Extensive experiments on both retrieval tasks (*i.e.*, zero-shot composed image retrieval, visual storytelling image retrieval and visual dialog image retrieval) and text generation tasks (*i.e.*, visual question answering) demonstrate the effectiveness of the proposed method. The code is available at: `https://github.com/dhg-wei/MCL`.

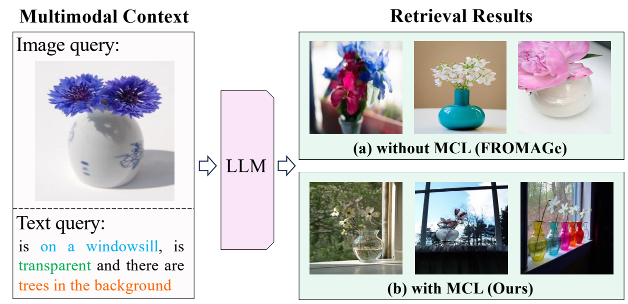


Figure 1. Comparison between FROMAGe (Koh et al., 2023b) and our MCL on zero-shot composed image retrieval. MCL enables the frozen LLM to retrieve accurate images that match the multimodal context (image and text queries).

1. Introduction

Recent research (Merullo et al., 2022; Li et al., 2023; Koh et al., 2023b;a) has shown that frozen Large Language Models (LLMs) can comprehend visual inputs and generate visual representations by learning a simple vision-language mapping through the utilization of image-text pairs. Aided by the strong contextual comprehension inherent in LLMs, Multimodal Large Language Models (MLLMs) demonstrate remarkable zero-shot abilities in multimodal tasks. Although trained on image captioning or image-text retrieval, these models excel in activities such as visual question answering, contextual image retrieval, and multimodal dialogue. This versatility showcases their broad applicability beyond their initial training focus. However, these methods that utilize image captioning and image-text retrieval tasks for vision-language mapping, mainly serving as 'modality translation', exhibit a deficiency in fostering sufficient cross-modal interaction. This limitation results in subpar performance on complex multimodal benchmarks (shown in Fig. 1), like zero-shot composed image retrieval, which require a profound understanding of multimodal contexts.

In this paper, we introduce a Multimodal Composition Learning (MCL) method to enhance the mapping between the vision and language modalities for MLLMs. One notable obstacle with this learning approach is the data intensiveness problem. Existing multimodal composition datasets consisting of an image query, a text query and a composed image target, which heavily rely on human labeling (Liu et al., 2021a; Wu et al., 2021). Relying on manual annotation limits the scope of these datasets to particular domains and poses challenges for scaling up, thus impeding the development of a comprehensive vision-language mapping. To obtain large-scale data for multimodal composition learning, we propose to leverage LLMs to enhance the existing web-

Part of this work was done when Wei Li was an Intern at National University of Singapore. [1]ReLER, CCAI, School of Computer Science and Technology, Zhejiang University, China. [2]School of Computing, National University of Singapore, Singapore. Correspondence to: Hehe Fan <hehefan@zju.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

collected image-caption pairs, resulting in a generated MultiModal Composition (MMC) dataset. Specifically, given a web-collected ⟨ref image, ref caption⟩ pair, we input the ref caption to LLM and prompt it to generate a text condition and a corresponding target caption, yielding ⟨ref image, ref caption, text condition, target caption⟩ tuples. For example, as shown in Fig. 2, given a ref image of 'an orange kitten' and the corresponding ref caption of 'cute orange kitten looking up', we first randomly generate a text condition of 'with a toy mouse', and then compose the target caption of 'a cute orange kitten playing with a toy mouse'. Note that we do not pursue obtaining the corresponding image of the target caption for training. Instead, we utilize the CLIP feature of the target caption as the visual supervision.

With the generated MMC dataset, we employ the proposed MCL method to bolster the bidirectional mapping between the visual space and language space. Specifically, we introduce two tasks: Multimodal-Context Captioning (MC-Cap) and Multimodal-Context Retrieval (MC-Ret). These tasks are designed to facilitate the learning of mappings from visual features to language space and to improve the extraction of visual representations from the multimodal context of LLM's input. Unlike traditional image captioning and image-text retrieval training objectives that focus on translating between the vision and language modalities, our proposed MCL approach aims to augment the model's proficiency in understanding and leveraging multimodal information. This is achieved by training the model to comprehend and utilize the image and language information, which is then used to generate targeted textual or visual representations. Our main contributions include:

- We propose a Multimodal Composition Learning (MCL) method for vision-language mapping. MCL can effectively enable a frozen LLM to perform accurate image retrieval and text generation within various multimodal contexts.
- We propose a MultiModal Composition (MMC) dataset, constructed by automatically augmenting existing web-collected image-text pairs. MMC contains 2.7 million tuples of ⟨ref image, ref caption, text condition, target caption⟩ .
- We propose a stacking retrieval mechanism to extract diverse multimodal information from LLM's multimodal context.
- Extensive experiments show the effectiveness of MCL on four zero-shot multimodal context understanding tasks, including composed image retrieval, visual storytelling image retrieval, visual dialog image retrieval and visual question answering.

2. Related Work

Vision-Language Mapping. In recent work, numerous efforts (Mokady et al., 2021; Tsimpoukelli et al., 2021; Merullo et al., 2022; Eichenberg et al., 2021; Li et al., 2023; Alayrac et al., 2022; Zhang et al., 2024; Yang et al., 2024) have been made to integrate the visual modality with Large Language Models (LLMs) by translating visual features into the frozen LLM space through the task of image captioning. Leveraging the robust textual capabilities of LLMs, these models are capable of performing traditional vision-language generation tasks such as image captioning and visual question answering. Furthermore, they are also extended to handle more sophisticated applications including visual dialogue and visual storytelling. Another research direction (Koh et al., 2023a;b) investigates the reverse process: mapping LLM representations into visual feature spaces. This is achieved by mapping the hidden states through learnable retrieval tokens to the CLIP (Radford et al., 2021) feature space, specifically for image-text retrieval tasks. In this paper, we further refine the vision-language mapping by introducing tasks based on multimodal composition. Unlike previous approaches that focused on mapping from one modality to another through tasks like image captioning and image-text retrieval, our proposed multimodal composition tasks require LLMs to synthesize information from various modalities, thereby enhancing the models' understanding and utilization of multimodal contexts.

Composed Image Retrieval. Composed Image Retrieval (CIR) aims to retrieve a target image based on multimodal queries that include a reference image and a text condition. Previous research (Baldrati et al., 2022a;b; Delmas et al., 2022; Lee et al., 2021; Liu et al., 2021b) primarily utilized human-labeled triplets for supervised training. Recent studies have explored performing the CIR task without the need for these human-labeled triplets. Pic2Word (Saito et al., 2023), CIRCO (Baldrati et al., 2023), and KEDs (Suo et al., 2024) map input images to pseudo text tokens, enabling the composition of image and text queries using the CLIP text encoder. Another line of research (Vaze et al., 2023; Gu et al., 2023; Liu et al., 2023b) has developed methods to automatically construct CIR triplets from image-caption pairs for training purposes. In this work, we execute zero-shot CIR using LLMs. We demonstrate that multimodal queries can be effectively synthesized within the LLM space.

Multimodel Data Augmentation with LLMs. Recent work (Liu et al., 2023a; Fan et al., 2023; Brooks et al., 2023; Zhang et al., 2023; Liu et al., 2023b; Li et al., 2024) uses LLMs for multimodal data refinement, enhancement, and extension. Brooks et al. (2023); Zhang et al. (2023); Liu et al. (2023b) aim to generate paired triplets (ref image, editing instructions, target image) for specific downstream tasks. In this paper, we leverage LLM, *i.e.*, Llama, to augment image-caption pairs into (ref image, text condition, target caption) pairs for MCL. Our work aims to introduce multimodal composition learning into vision-language alignment to enhance the multimodal context understanding capability.

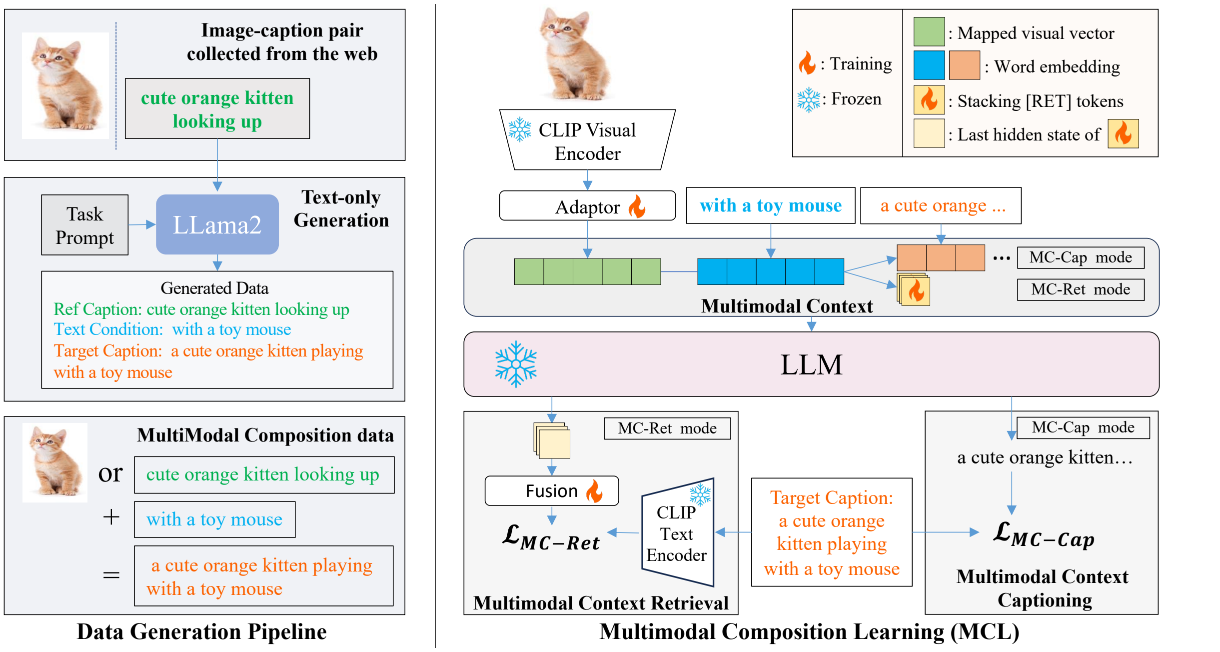


Figure 2. Framework of MCL. Given ⟨image, caption⟩ pairs, we feed the caption along with a task prompt to Llama2 to generate both a text condition and a target caption. By integrating the initial image with the generated text condition and target caption, we facilitate multimodal composition learning, enabling a frozen language model to compose multimodal inputs and output visual representations.

3. Method

This section delineates the proposed MCL framework where an overview is illustrated in Figure 2. Section 3.1 details the data generation process of our MMC dataset. In Section 3.2, we introduce how to enable a frozen LLM to process visual input. Section 3.3 describes the process to extract visual representation from LLM's multimodal context.

3.1. Data Generation

Collecting paired data (*i.e.*, a multimodal query and a composed target) for multimodal composition learning poses significant challenges. This lead to the unavailability of large-scale training data and limits the development of multimodal composition learning. To address this, we propose to automatically generate a large-scale multimodal composition dataset from existing ⟨image, caption⟩ pairs by leveraging an off-the-shelf LLM (*i.e.*, Llama2). Given a reference image I_{ref} and its associated caption T_{refc}, we input the T_{refc} into an LLM along with a task-specific prompt. The LLM then generates a free-text condition T_{con}, which could serve as an editing order to alter attributes and objects, or describe the differences between the reference image and the target image. Following this, the LLM is tasked with generating a target caption T_{tgtc} by composing the reference caption T_{refc} and the newly generated text condition T_{con}. As a result, we derive a $\langle I_{\text{ref}}, T_{\text{refc}}, T_{\text{con}}, T_{\text{tgtc}}\rangle$ tuple from a $\langle I_{\text{ref}}, T_{\text{refc}}\rangle$ pair, which can be automatically collected from the web.

3.2. Mapping Visual Input to LLMs

Adapting Frozen LLMs to Visual Input. Following the latest advancements in vision-Language research (Mokady et al., 2021; Merullo et al., 2022; Koh et al., 2023b), we employ a linear mapping layer (*i.e.*, an adaptor) that maps CLIP visual features into the LLM's embedding space. Given an input image $\mathbf{I}$, we first utilize a frozen CLIP visual encoder E_{image} to extract the corresponding visual feature. Subsequently, a linear mapping layer f_{map} is applied to map this visual feature into the LLM's embedding space, resulting in n visual vectors $\mathbf{V} = [\boldsymbol{v_0}, \boldsymbol{v_1}, ..., \boldsymbol{v_n}] = f_{\text{map}}(E_{\text{image}}(\boldsymbol{I}))$. The visual vector's dimension matches the LLM's word embedding dimension.

Naive Image Captioning Objective. Previous methods employ a conventional image captioning objective to train the mapping layer f_{map} by predicting the next token conditioned on both the visual tokens and the previous caption tokens. The objective can be formulated as:

$$\mathcal{L}_{\text{Cap}}(\theta_m) = -\frac{1}{|t|}\sum_{i=1}^{|t|} \log P\Big(t_i|\mathbf{V}, t_{<i}\Big), \qquad (1)$$

where t_i represents the i_{th} caption token, θ_m denotes the weight of mapping layer f_{map} and P denotes a frozen LLM.

Multimodal-Context Captioning (MC-Cap) Objective. In this work, we enhance this vision-to-language mapping by incorporating it with the generated MMC dataset described in Section 3.1. Given a triplet $\langle I_{\text{ref}}, T_{\text{con}}, T_{\text{tgtc}} \rangle$, the LLM is tasked to predict the next token conditioned on the visual vectors $\mathbf{V}$, text conditions tokens and previous target caption tokens. The objective is described as:

$$\mathcal{L}_{\text{MC-Cap}}(\theta_m) = -\frac{1}{|t|} \sum_{i=1}^{|t|} \log P\Big(t_i | \mathbf{V}, c_1, .., c_{|c|}, t_{<i}\Big), \tag{2}$$

where c_i denotes the i_{th} token of text condition and t_i denotes the i_{th} token of target caption.

Compared to the conventional image captioning objective (*i.e.*, Equation 1), our proposed objective offers distinct advantages. **Enhanced Linguistic Visual Mapping**: Our training objective refines the mapping process by incorporating textual cues (*i.e.*, the text condition T_{con}), leading to a textual-aware mapping that effectively integrates visual features into the language model's semantic space. **Enhanced Textual Interaction**: The language model is tasked to query the mapped visual vectors based on the text condition to derive the target caption. This ensures the mapped visual vectors are optimized to support textual queries.

3.3. Extracting Visual Representations from LLMs

In this section, we delineate how to extract visual representation from LLM's representation space. Following (Koh et al., 2023b), we leverage learnable tokens to extract visual information from the LLM multimodal context. Specifically, a special token `[RET]` is appended following the context tokens. The `[RET]` token is used to prompt the LLM to gather visual information from the multimodal context. The last hidden state of the `[RET]` token is used to output the corresponding visual representation.

Naive Image-Text Retrieval Objective. FROMAGe (Koh et al., 2023b) leverages an image-text retrieval task to train the `[RET]` embedding. Specifically, given a paired image and caption, they append a `[RET]` token after the caption tokens as input to the LLM. The last hidden state of `[RET]` token is used as the LLM's output, represented as $h(\texttt{[RET]}|\mathbf{T})$, where $\mathbf{T}$ denotes the caption tokens. $h(\texttt{[RET]}|\mathbf{T})$ is then projected to CLIP latent space through a simple linear layer, represented as $p_v = f_{\text{proj}}(h(\texttt{[RET]}|\mathbf{T}))$. An infoNCE (Oord et al., 2018) loss is employed to align the projected embedding and the CLIP visual feature of the target image. The objective is formulated as:

$$\begin{aligned} &\mathcal{L}_{\text{Ret}}(\texttt{[RET]}, \theta_p) \\ &= -\frac{1}{N} \sum_{i=1}^{N} \left(\log \frac{\exp(\text{sim}(p_v, \boldsymbol{e}_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(p_v, \boldsymbol{e}_j)/\tau)} \right), \end{aligned} \tag{3}$$

where θ_p denotes the weight of project layer f_{proj}, $\text{sim}(\cdot, \cdot)$ denotes cosine similarity function, and $\boldsymbol{e}_i = E_{\text{image}}(\boldsymbol{I}_i)$.

Multimodal-Context Retrieval (MC-Ret) Objective. Trained with naive image-text matching objective, `[RET]` bridges the LLM context with the CLIP feature space. However, in this case, the `[RET]` token primarily functions as a text summarization token, condensing the text context into CLIP feature space, lacking the ability to selectively extract target information based on the given multimodal context. To this end, we introduce a MC-Ret objective to enhance the capability of extracting information from multimodal context. Given triplet $\langle I_{\text{ref}}, T_{\text{con}}, T_{\text{tgtc}} \rangle$, we input the I_{ref} and T_{con} to the LLM with `[RET]` token attached at the end. In this scenario, `[RET]` learns to compose the multimodal context to match the target caption T_{tgtc}. The objective can be formulated as:

$$\begin{aligned} &\mathcal{L}_{\text{MC-Ret}}(\texttt{[RET]}, \theta_p, \theta_m) \\ &= -\frac{1}{N} \sum_{i=1}^{N} \left(\log \frac{\exp(\text{sim}(p_v, \boldsymbol{e}_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(p_v, \boldsymbol{e}_j)/\tau)} \right), \end{aligned} \tag{4}$$

where $p_v = f_{\text{proj}}(h(\texttt{[RET]}|\mathbf{V}, \mathbf{T}))$, $\mathbf{V}$ denotes the mapped visual features of I_{ref}, $\mathbf{T}$ denotes the caption tokens of T_{con} and $\mathbf{e}$ denotes the CLIP text feature of T_{tgtc}.

The MC-Ret objective brings the following benefits: (a) The `[RET]` token is trained within multimodal contexts, making it better adapted to handle multimodal inputs. (b) The `[RET]` token learns to selectively extract information based on the multimodal context, rather than indiscriminately condensing all the input.

Multiple Retrieval Tokens with Sequential Order. A straightforward and effective method to enhance the visual information extraction is to append more `[RET]` tokens after the context tokens in a sequential order. We adapt Equation 3 to multiple `[RET]` tokens scenario by simply modifying the output feature p_v as follows:

$$\begin{aligned} p_v &= f_{\text{fusion}}(h_1, \ldots, h_r), \\ h_i &= h(\texttt{[RET]}_i \mid \mathbf{V}, \mathbf{T}, \texttt{[RET]}_{<i}), \end{aligned} \tag{5}$$

where r represents the number of `[RET]` token and f_{fusion} denotes a fusion function that integrates multiple hidden states into a single vector. The multiple `[RET]` tokens are excepted to extract diverse information from the multimodal context. However, we find that the multiple `[RET]` tokens sometimes perform worse than the single `[RET]` token. This phenomenon can be attributed to the fact that the hidden state of $\texttt{[RET]}_i$ is significantly influenced by the preceding $\texttt{[RET]}_{<i}$ tokens due to the intrinsic properties of LLMs. Consequently, the adjacent `[RET]` tokens tends to focus on similar content. This tendency contradicts our goal of extracting diverse information from the context.

Stacking Retrieval Mechanism. To mitigate this issue and extract diverse information from the LLM context, we introduce a stacking retrieval mechanism. In this approach, multiple [RET] tokens are appended after the context tokens, arranged in a stacking order instead of a traditional sequential order. The output feature p_v is represented as:

$$p_v = f_{\text{fusion}}(h_1, ..., h_r), \quad h_i = h(\texttt{[RET]}_i \mid \mathbf{V}, \mathbf{T}) \quad (6)$$

In this case, the output of each [RET] token is conditioned only by the multimodal context, independent of other [RET] tokens. The stacking approach allows [RET] tokens to extract more diverse information from the LLM context, rather than concentrating on similar content. We implement the stacking mechanism by adding extra attention masks between the RET tokens.

3.4. Model Training

We combine the four aforementioned objectives for vision-LLM mapping. $\mathcal{L}_{\text{Cap}}$ and $\mathcal{L}_{\text{Ret}}$ are based on the $\langle\textsf{ref image}, \textsf{ref caption}\rangle$ pairs, while the proposed $\mathcal{L}_{\text{MC-Cap}}$ and $\mathcal{L}_{\text{MC-Ret}}$ are based on the triplets $\langle\textsf{ref image}, \textsf{text condition}, \textsf{target caption}\rangle$. The combined loss function is expressed as:

$$\mathcal{L} = \lambda_{\text{Cap}}(\mathcal{L}_{\text{Cap}} + \mathcal{L}_{\text{MC-Cap}}) + \lambda_{\text{Ret}}(\mathcal{L}_{\text{Ret}} + \mathcal{L}_{\text{MC-Ret}}),$$

where λ_{Cap} and λ_{Ret} denote the weights of generation losses and retrieval losses, respectively.

Implementation Details. We employ CLIP ViT-L/14 as our image-text retrieval model. We utilize OPT-2.7B, OPT-6.7B and Llama2-7B as the LLM backbone. The input image is mapped to 4 visual vectors in LLM space. The number of [RET] tokens is set to 5. We adopt a two-layer transformer with a mean pooling as the fusion function for multiple [RET] tokens. MCL is trained on MMC for 50,000 iterations with a batchsize of 64. Both the LLM and CLIP model are frozen. The loss weights λ_{Cap} and λ_{Ret} in Equation 7 is set to 0.5 and 1.0 respectively. The temperature τ in Equation 3 and Equation 4 is set to 0.07.

4. Experiments

As shown in Figure 3, our proposed MCL effectively enables the model to perform multimodal tasks within arbitrary multimodal input. In this section, we first conduct extensive experiments on conventional multimodal image retrieval tasks, namely zero-shot composed image retrieval (in Section 4.1). Furthermore, we conduct experiments on dense multimodal context understanding tasks where the input encompasses multiple images and texts (in Section 4.2). Then, to assess the multimodal understanding capability in text generation tasks, we conduct experiments on visual question answering (in Section 4.3). Ablation studies and analysis are in Section 4.4.

4.1. Composed Image Retrieval

Benchmarks and Metrics. We evaluate MCL on three zero-shot CIR benchmarks: CIRCO (Baldrati et al., 2023), CIRR (Liu et al., 2021a) and GeneCIS (Vaze et al., 2023). CIRCO is an open-domain zero-shot CIR benchmark with multiple annotated ground truths. Following existing methods, we report the fine-grained metric of mean Average Precision (mAP@K) on CIRCO. The mAP@K metrics are computed considering all the ground truth images for each query. For CIRR and GeneCIS, we report the Recall@K metric.

Baselines and Competing Methods. We compare our approach with several baselines and recent zero-shot CIR methods in the zero-shot setting, including: (1) ***Image-only***: The CLIP visual feature of the reference image is used to retrieve the target image. (2) ***Text-only***: The CLIP text feature of the text condition is used to retrieve the target image. (3) ***Image+Text***: The CLIP visual feature of the reference image and the CLIP text feature of the text condition are summed together to retrieve the target image. (4) CLIP-based textual inversion methods: ***Pic2Word*** (Saito et al., 2023) and ***SEARLE*** (Baldrati et al., 2023). (5) ***CompoDiff*** (Gu et al., 2023): Combiner (Baldrati et al., 2022a) trained on generated triplets (6) Combiner-MMC trained on our proposed MMC. (7) LLM-based approaches: ***FROMAGe*** (Koh et al., 2023b).

Results and Analysis. Table 1 shows the results on the zero-shot CIR tasks. Overall, MCL shows impressive results on three benchmarks, outperforming previous CLIP-based zero-shot composing methods and LLM-based methods. We can draw a few conclusions from the table: **(1) The proposed multimodal composition learning method significantly improves the MLLM's capability in composing multimodal context.** The FROMAGe MLLM which is trained on the modality translation tasks (*i.e.*, image captioning and image-text retrieval) performs only better than the single modality baselines (*i.e.*, image-only and text-only). It indicates that the modality translation tasks are not enough to enable the LLM to compose multimodal contexts. Benefiting from the multimodal composition training, MCL effectively composes the multimodal context and extracts the target representation for retrieval. **(2) LLM is better than the CLIP text encoder for multimodal composition.** MCL largely outperforms previous zero-shot CIR approaches, such as Pic2Word (Mokady et al., 2021) and SEARLE (Baldrati et al., 2023), that compose the multimodal input in the text encoder of CLIP. It indicates that the frozen LLM space can compose the image and text inputs, even though the LLM is pre-trained on text corpus. The CLIP text encoder, which is trained through image-text contrastive learning, usually faces challenges in comprehending object relations, word order, and logical structures (Yuksekgonul et al., 2022; Ma et al., 2023; Thrush et al., 2022;

Table 1. Results on zero-shot CIR benchmarks. The best and second-best scores are highlighted in **bold** and underlined, respectively.

Method	LLM	CIRCO (mAP@K)				CIRR				GeneCIS
		K=5	K=10	K=25	K=50	R@1	R@5	R@50	R_s@1	R@1 (avg)
Image-only	Non-LLM	2.79	3.18	3.75	4.12	7.13	23.04	56.63	20.55	11.0
Text-only	Non-LLM	2.50	2.64	3.11	3.38	20.55	44.17	78.94	<u>60.74</u>	9.1
Image + Text	Non-LLM	6.37	7.04	8.11	8.72	12.27	35.81	77.04	33.33	12.6
Pic2Word	Non-LLM	8.72	9.51	10.64	11.29	23.90	51.70	87.80	54.12	11.2
SEARLE	Non-LLM	11.68	12.73	14.33	15.12	<u>24.22</u>	52.41	88.63	53.71	12.3
ComposDiff	Non-LLM	12.55	13.36	15.83	16.43	18.24	53.14	90.25	57.42	14.9
Combiner-MMC	Non-LLM	13.22	14.07	15.53	16.32	21.74	51.54	88.48	49.27	14.0
FROMAGe	OPT-6.7B	4.00	4.44	5.26	5.73	10.96	31.40	72.97	34.07	14.3
MCL (ours)	OPT-2.7B	14.55	15.79	17.38	18.27	23.28	54.17	90.05	58.24	15.8
MCL (ours)	OPT-6.7B	<u>15.14</u>	<u>16.13</u>	<u>17.88</u>	<u>18.82</u>	24.15	<u>55.98</u>	<u>90.92</u>	59.52	<u>16.1</u>
MCL (ours)	Llama2-7B	**17.67**	**18.86**	**20.80**	**21.68**	**26.22**	**56.84**	**91.35**	**61.45**	**16.3**

Wang et al., 2023). This limitation constrains its capabilities in multimodal composition. LLM, in contrast, exhibits a proficient ability to easily comprehend these complex expressions. We provide more qualitative comparisons and analysis of the logical word understanding in Appendix B.1. **(3) MCL benefits from stronger LLM.** Integrating MCL with more advanced LLMs results in uniform enhancements across three benchmarks. These improvements stem from an enhanced representation space and a superior ability to understand context. Further qualitative results and analysis detailing the effects of various LLM backbones are available in the Appendix B.2. **(4) MMC can be used for conventional CIR training.** We use the MMC dataset to train a Combiner (Baldrati et al., 2022a) model, which is a classic CIR method that employs a simple combiner component to integrate features from the image encoder and text encoder of CLIP. The Combiner model trained on the MMC dataset achieves competitive results, approaching previous textual inversion-based method SEARlE, demonstrating the effectiveness of the generated MMC dataset. Despite being trained on the same MMC dataset, a substantial performance gap exists between the Combiner and the proposed MCL, underscoring the efficacy of LLM in tasks involving multimodal composition.

4.2. Dense Multimodal Context Understanding

To investigate MCL's multimodal understanding ability in more complex scenarios, we consider the dense multimodal context understanding tasks where the input encompasses multiple images and texts. Conventional image-text matching models are restricted to performing retrieval between a single image and a single text. Similarly, composed image retrieval models are also constrained to a single reference image and a single text query. Thanks to the LLM, our approach, despite not being explicitly trained on data with multiple images and texts, showcases its adeptness in understanding the dense multimodal context.

Table 2. Zero-shot image retrieval results on Visual Storytelling. † indicates input images from the current story sequence are masked in the retrieval gallery.

Method	Inputs	R@1	R@5	R@10
CLIP ViT-L/14	1 caption	**11.9**	**25.5**	32.2
FROMAGe (OPT-6.7B)	1 caption	11.3	24.6	32.1
MCL (OPT-2.7B)	1 caption	8.6	20.9	28.5
MCL (OPT-6.7B)	1 caption	9.4	22.1	29.3
MCL (Llama2-7B)	1 caption	11.4	25.8	**33.9**
CLIP ViT-L/14	5 captions	5.9	19.5	28.0
FROMAGe (OPT-6.7B)	5 captions	10.8	23.8	31.7
MCL (OPT-2.7B)	5 captions	9.8	25.2	35.7
MCL (OPT-6.7B)	5 captions	11.9	28.8	38.4
MCL (Llama2-7B)	5 captions	**13.7**	**32.9**	**42.7**
CLIP ViT-L/14	5 captions, 4 images†	2.4	21.3	34.0
FROMAGe (OPT-6.7B)	5 captions, 4 images†	18.2	42.7	51.8
GILL (OPT-6.7B)	5 captions, 4 images†	20.3	45.0	53.7
MCL (OPT-2.7B)	5 captions, 4 images†	21.8	44.6	53.9
MCL (OPT-6.7B)	5 captions, 4 images†	22.5	46.5	55.8
MCL (Llama2-7B)	5 captions, 4 images†	**23.1**	**46.7**	**56.1**

Following Koh et al. (2023b), we conduct zero-shot experiments on Visual Storytelling (Huang et al., 2016) and Visual Dialogue (Das et al., 2017). Unlike CIR tasks, which involve only a single reference image and a single textual condition, both Visual Dialogue and Visual Storytelling entail long and intricate contexts. In these more challenging scenarios, the model is required to not only understand the multimodal context but also to efficiently extract crucial information from multimodal dialogues or narratives.

Visual Storytelling Results. Each example in the visual storytelling dataset comprises five temporally ordered image-text pairs, we report Recall@K of the last image as the metric. Following (Koh et al., 2023b), we explore several experimental settings featuring different input configurations: (1) single last caption as input; (2) input consisting of all five captions; (3) input incorporating five captions along with four associated images. Table 2 shows the results. We

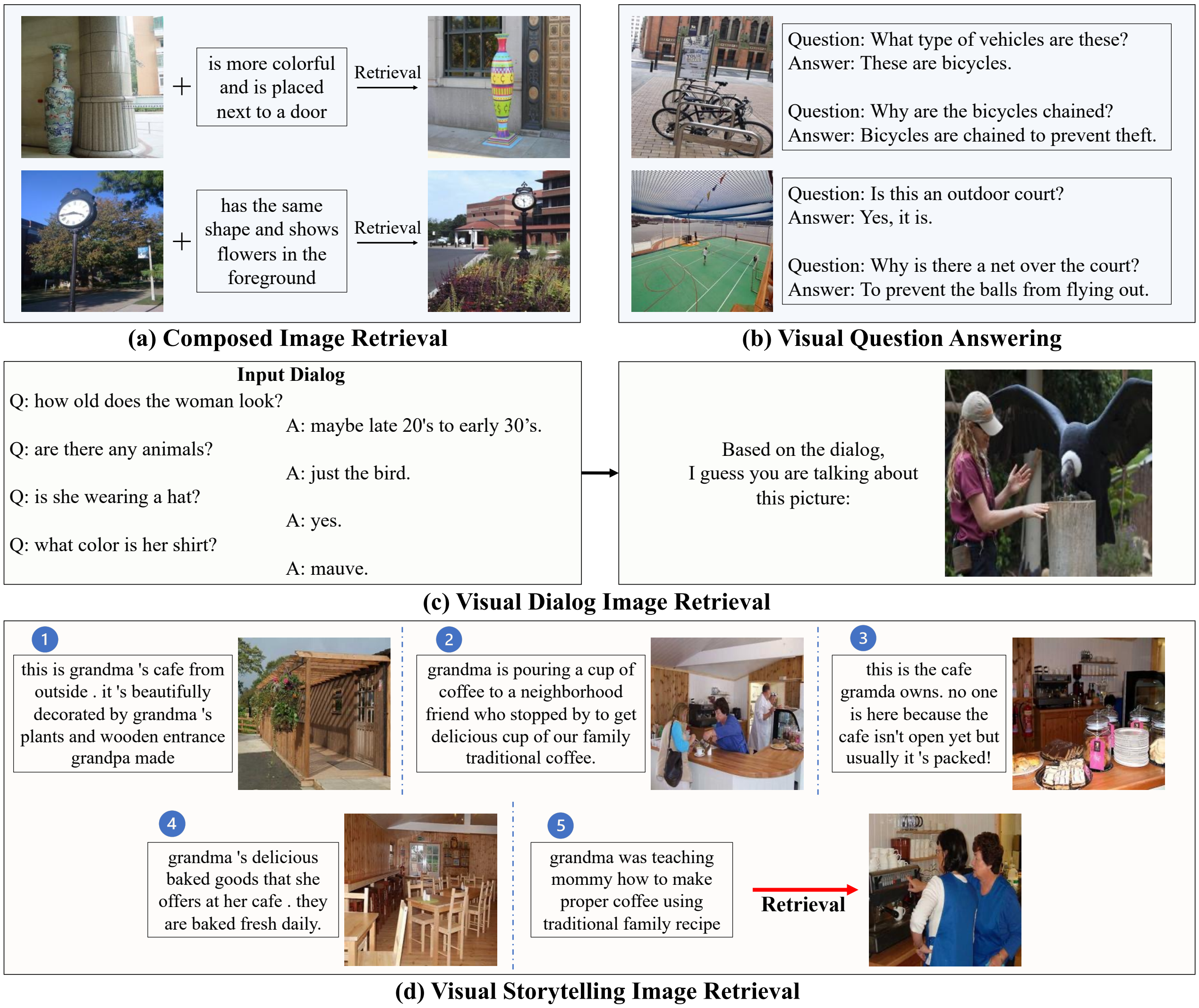


Figure 3. Examples of multimodal understanding tasks with different types of multimodal context inputs. MCL is capable of processing situations where there is a single image and a single text condition input (*e.g.*, composed image retrieval and visual question answering). Furthermore, it can also adjust to image retrieval tasks that require multiple continuous inputs (*e.g.*, visual dialogue image retrieval and visual storytelling image retrieval).

can draw the following main conclusions: **(a) MCL benefits from dense contexts.** When the input context is increased from '1 caption' to '5 captions', MCL shows significant improvements. For instance, the R@1 score of the Llama2 model increased from 11.4% to 13.7%. In contrast, the performance of the CLIP model decreases from 11.9% to 5.9%. Similarly, the LLM-based method FROMAGe shows a slight decrease with a richer context. These results suggest that our proposed multimodal composition learning effectively enhances the capability to extract information from dense contexts. **(b) MCL benefits from multimodal context.** As the input incorporates five captions along with four associated images, the performance of MCL further improves, outperforming previous LLM-based approaches FROMAGe (Koh et al., 2023b) and GILL (Koh et al., 2023a). This indicates that MCL is effective in extracting information from a multimodal context. We provide more qualitative results on Visual Storytelling in Appendix B.3.

Visual Dialog Results. Each sample in Visual Dialog contains one image and a conversation about this image. We take the conversation as the input to retrieve the corresponding image. The results are shown in Table 3. The proposed MCL outperforms CLIP baseline and prior LLM-based retrieval methods by a large margin. This demonstrates MCL's capability to extract visual representations not only from simple caption-style context but also from intricate dialog-style context.

Table 3. Zero-shot image retrieval results on Visual Dialog.

Method	R@1	R@5	R@10
CLIP ViT-L/14	17.7	38.9	50.2
FROMAGe (OPT-6.7B)	20.8	44.9	56.0
MCL (OPT-2.7B)	25.6	51.9	65.2
MCL (OPT-6.7B)	27.2	51.0	64.0
MCL (Llama2-7B)	**29.8**	**57.1**	**69.4**

4.3. Visual Question Answering

To further explore MCL's multimodal capability, we conduct experiment on VQAv2 (Goyal et al., 2017), which requires the model to generate the answer based on the $\langle \mathsf{image}, \mathsf{question} \rangle$ pair. The results are shown in Table 4. We use the prompt *"Question:{Question} Answer:for this image, the answer is {Answer}"* for evaluation. We find that such a prompt effectively prevents the LLM from generating unrelated content. MCL outperforms similar parameter-efficient methods which are trained with naive image captioning objectives and image-text retrieval objectives. This result demonstrates that our proposed composition learning effectively integrates the visual feature into the LLM space thereby providing advantages for a range of multimodal tasks. We note that these zero-shot results are lower than recent SOTA MLLMs (Li et al., 2023; Zhu et al., 2023; Alayrac et al., 2022; Ye et al., 2023), as they are trained with significantly more computing and data, especially some of them employ in-domain data for training (*i.e.*, the MSCOCO (Lin et al., 2014) dataset, has the same data source with VQAv2).

Table 4. Zero-shot results on VQAv2 val set. † denotes reproduced results with our prompts.

Model	LLM	Acc@zero-shot
Frozen (Tsimpoukelli et al., 2021)	GPT-2	29.5
MAGMA (Eichenberg et al., 2021)	GPT-J-6B	32.7
LinearMapping (Merullo et al., 2022)	GPT-J-6B	33.3
Fromage (Koh et al., 2023b)†	OPT-6.7B	36.8
GILL (Koh et al., 2023a)†	OPT-6.7B	38.8
MCL (Ours)	OPT-2.7B	38.4
MCL (Ours)	OPT-6.7B	40.2
MCL (Ours)	Llama2-7B	**42.6**

4.4. Analysis and Ablations

Understand MCL by Visualizing the Relevance between Context Tokens and Retrieval Tokens. In Figure 4, we visualize the relevance between the context tokens and `[RET]` tokens. The relevance score is calculated by aggregating the token relevance across the attention layers as described in (Chefer et al., 2021). From the figure we can find that: (a) Our composition learning enables the model to effectively compose the visual and textual input to accurately retrieve the target image. For instance, in the first example, the model retrieves the target by identifying cues like 'same color', 'congested street', and 'stopped'. Conversely, the model without composition learning tends to concentrate on visual tokens or with only partial textual cues, leading to incorrect retrieval results. (b) The proposed stacking retrieval mechanism makes the learned retrieval tokens focus on different contexts. For instance, in the third example, the retrieval tokens tend to focus on 'different color', 'seen in the side' and 'sky in background'. Conversely, the naive sequential retrieval tokens mostly focus on the 'different color', 'seen in the side' and overlook the 'sky in the background', leading to incorrect results.

Table 5. Results for the ablation study on the proposed Stacking Retrieval (S.R.) mechanism and MCL objectives, respecrtively. The CIRCO test set is used for evaluation.

Method	$\mathcal{L}_{\text{Cap}}$	$\mathcal{L}_{\text{Ret}}$	$\mathcal{L}_{\text{MC-Cap}}$	$\mathcal{L}_{\text{MC-Ret}}$	mAP@5	mAP@10	mAP@25	mAP@50
Single `[RET]` token	✓	✓	✓	✓	17.07	17.87	19.65	20.62
5 `[RET]` tokens w/o S.R.	✓	✓	✓	✓	16.48	17.67	19.48	20.35
5 `[RET]` tokens w/ S.R.	✓	✓	✓	✓	17.67	18.86	20.80	21.68
Naive Mapping	✓	✓			4.38	4.70	5.44	5.85
Naive + MC-Cap	✓	✓	✓		6.58	6.97	7.82	8.36
Naive + MC-Ret	✓	✓		✓	16.73	17.80	19.58	20.51
MC-Cap + MC-Ret			✓	✓	17.55	18.67	20.63	21.62
MCL	✓	✓	✓	✓	17.67	18.86	20.80	21.68

Ablations on Stacking Retrieval Mechanism. Table 5 shows the results of the ablation study on Stacking Retrieval Mechanism. When the number of the `[RET]` token increase to five, the performance decreases. This decline is attributable to the inherent characteristics of LLMs, where adjacent tokens heavily influence one another, leading to focusing on similar contexts as shown in Figure 4. The stacking retrieval mechanism allows the multiple `[RET]` tokens to extract diverse information from multimodal contexts.

Multimodal Composition Learning Loss. Table 5 shows the results of ablation study on the proposed objectives. It shows that our MCL objectives (*i.e.*, MC-Cap and MC-Ret) significantly improve the performance on composed image retrieval tasks. MC-Ret objective is most effective because it jointly optimizes both the visual input and visual output mappings. The model works well only with the proposed MC-Cap and MC-Ret objectives. Adding naive image captioning and image-text retrieval objectives results in a slight improvement.

5. Conclusion

In this paper, we propose multimodal composition learning for vision-language mapping. Compared to previous MLLMs trained with image captioning and image-text retrieval tasks, our MCL shows superior performance in various multimodal scenarios. We hope MCL could inspire future exploration of aligning LLMs with other modalities.

Impact Statement

This work primarily focuses on improving the functionality and efficiency of multimodal composition tasks. Our advancements in this field aim to enrich the interaction

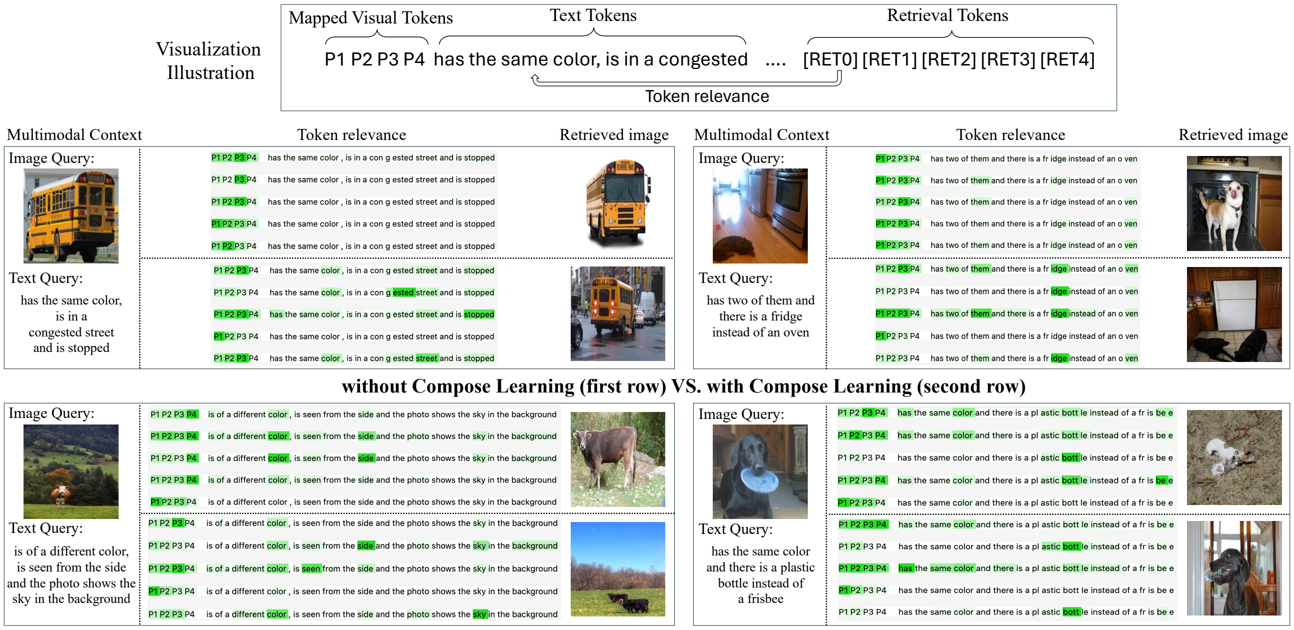


Figure 4. Visualization of the relevance between the input multimodal context tokens and learned retrieval tokens. The deeper the green color, the higher the relevance. P1, P2, P3 and P4 denote the mapped visual tokens. (a) Our composition learning enables the model to effectively compose the visual and textual input to accurately retrieve the target image. (b) The proposed stacking retrieval mechanism makes the learned retrieval tokens focus on different contexts.

between humans and technology, enhancing access to mutlimodal information. These improvements can lead to more effective and user-friendly search engines, which are integral to various aspects of modern life. Furthermore, by addressing the limitations of existing systems, we contribute to a more accurate and efficient processing of multimodal data, avoiding potential misinformation or misinterpretation that can arise from less sophisticated models.

Acknowledgements

This work was supported by National Key R&D Program of China (No. 2023YFC3305600), the National Natural Science Foundation of China (U2336212) and the Fundamental Research Funds for the Central Universities (No. 226-2022-00051), Lu's Graduate Education International Exchange Foundation and the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. The computational work was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Baldrati, A., Bertini, M., Uricchio, T., and Del Bimbo, A. Conditioned and composed image retrieval combining and partially fine-tuning CLIP-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4959–4968, 2022a.

Baldrati, A., Bertini, M., Uricchio, T., and Del Bimbo, A. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21466–21474, 2022b.

Baldrati, A., Agnolucci, L., Bertini, M., and Del Bimbo, A. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference of Computer Vision*, 2023.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Chefer, H., Gur, S., and Wolf, L. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 397–406, 2021.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D., and Batra, D. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 326–335, 2017.

Delmas, G., de Rezende, R. S., Csurka, G., and Larlus, D. ARTEMIS: Attention-based retrieval with text-explicit matching and implicit similarity. *arXiv preprint arXiv:2203.08101*, 2022.

Eichenberg, C., Black, S., Weinbach, S., Parcalabescu, L., and Frank, A. Magma–multimodal augmentation of generative models through adapter-based finetuning. *arXiv preprint arXiv:2112.05253*, 2021.

Fan, L., Krishnan, D., Isola, P., Katabi, D., and Tian, Y. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36, 2023.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Gu, G., Chun, S., Kim, W., Jun, H., Kang, Y., and Yun, S. CompoDiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*, 2023.

Huang, T.-H., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C. L., Parikh, D., Vanderwende, L., Galley, M., and Mitchell, M. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pp. 1233–1239, 2016.

Koh, J. Y., Fried, D., and Salakhutdinov, R. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*, 2023a.

Koh, J. Y., Salakhutdinov, R., and Fried, D. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning*, 2023b.

Lee, S., Kim, D., and Han, B. CoSMo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 802–812, 2021.

Li, J., Li, D., Savarese, S., and Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023.

Li, W., Fan, H., Wong, Y., Kankanhalli, M., and Yang, Y. Topa: Extend large language models for video understanding via text-only pre-alignment. *arXiv preprint arXiv:2405.13911*, 2024.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2023a.

Liu, Y., Yao, J., Zhang, Y., Wang, Y., and Xie, W. Zero-shot composed text-image retrieval. *BMVC*, 2023b.

Liu, Z., Rodriguez-Opazo, C., Teney, D., and Gould, S. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2125–2134, 2021a.

Liu, Z., Rodriguez-Opazo, C., Teney, D., and Gould, S. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2125–2134, 2021b.

Ma, Z., Hong, J., Gul, M. O., Gandhi, M., Gao, I., and Krishna, R. CREPE: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10910–10921, 2023.

Merullo, J., Castricato, L., Eickhoff, C., and Pavlick, E. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022.

Mokady, R., Hertz, A., and Bermano, A. H. ClipCap: CLIP prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Saito, K., Sohn, K., Zhang, X., Li, C.-L., Lee, C.-Y., Saenko, K., and Pfister, T. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings* of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19305–19314, 2023.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. 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)*, pp. 2556–2565, 2018.

Suo, Y., Ma, F., Zhu, L., and Yang, Y. Knowledge-enhanced dual-stream zero-shot composed image retrieval. *arXiv preprint arXiv:2403.16005*, 2024.

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., and Ross, C. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tsimpoukelli, M., Menick, J. L., Cabi, S., Eslami, S., Vinyals, O., and Hill, F. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

Vaze, S., Carion, N., and Misra, I. GeneCIS: A benchmark for general conditional image similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6862–6872, 2023.

Wang, H., Li, Y., Yao, H., and Li, X. Clipn for zero-shot ood detection: Teaching clip to say no. *arXiv preprint arXiv:2308.12213*, 2023.

Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., and Feris, R. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11307–11317, 2021.

Yang, Z., Chen, G., Li, X., Wang, W., and Yang, Y. Doraemongpt: Toward understanding dynamic scenes with large language models (exemplified as a video agent). In *ICML*, 2024.

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., and Zou, J. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022.

Zhang, S., Yang, X., Feng, Y., Qin, C., Chen, C.-C., Yu, N., Chen, Z., Wang, H., Savarese, S., Ermon, S., et al. Hive: Harnessing human feedback for instructional visual editing. *arXiv preprint arXiv:2303.09618*, 2023.

Zhang, Y., Fan, H., and Yang, Y. Prompt-aware adapter: Towards learning adaptive visual tokens for multimodal large language models. *arXiv*, 2024.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A. Detailed Composed Image Retrieval Results and Analysis

Analysis on CIRR. Table 6 shows the detailed results on CIRR test set. MCL achieves SOTA on most metrics. Notably, the CIRR benchmark has a strong bias towards the text modality input (Saito et al., 2023; Baldrati et al., 2023). The Text-only baseline surpasses the Image+Text baseline a lot and even outperforms most zero-shot CIR methods on the $Recall_{\text{subset}}@K$ metrics. Despite this modality bias, MCL still achieves superior performance.

Table 6. Quantitative results on CIRR test set. Below the dashed line are LLM-based methods.

Method	Recall@K				Recall$_{\text{Subset}}$@K		
	$K=1$	$K=5$	$K=10$	$K=50$	$K=1$	$K=2$	$K=3$
Image-only	7.13	23.04	32.99	56.63	20.55	40.96	61.04
Text-only	20.55	44.17	55.95	78.94	60.74	80.38	90.72
Image+Text	12.27	35.81	48.48	77.04	33.33	57.78	75.95
TransAgg	25.04	53.98	67.59	88.94	55.33	76.82	88.94
CompoDiff	18.24	53.14	**70.82**	90.35	57.42	77.10	87.90
Pic2Word	23.90	51.70	65.30	87.80	54.12	74.63	87.61
SEARLE	24.22	52.41	66.29	88.63	53.71	74.63	87.61
Combiner-MMC	21.74	51.54	65.33	88.48	49.28	72.88	87.01
FROMAGe-(OPT-6.7B)	10.96	31.40	44.33	72.97	34.07	58.84	76.80
MCL-(OPT-2.7B)	23.28	54.17	67.16	90.05	58.24	79.37	90.51
MCL-(OPT-6.7B)	24.15	55.98	69.21	90.82	59.52	80.34	91.13
MCL-(LLama2-7B)	**26.22**	**56.84**	70.00	**91.35**	**61.45**	**81.61**	**91.93**

Analysis on GeneCIS. Table 7 shows the detailed results on GeneCIS. GeneCIS introduces four unique tasks: Focus Attribute, Change Attribute, Focus Object, and Change Object. For each task, only a single object name or attribute name is provided. This setup differs significantly from prior benchmarks such as CIRR and CIRCO, which often provide caption-style text conditions. Overall, MCL achieves superior performance on the Avg R@1 metric. In Focus Attribute task, FROMAGe achieves the best performance. This is because the image modality contains major information in the focus attribute task, where the Image-only baseline achieves 18.2% at R@1, outperforming the Image+Text baseline. For the other three balance tasks, MCL consistently achieves better performance, demonstrating its effectiveness in multimodal composition.

Table 7. Quantitative results on GeneCIS.

Method	Focus Attribute			Change Attribute			Focus Object			Change Object			
	R@1	R@ 2	R@3	R@1	R@ 2	R@3	R@1	R@ 2	R@3	R@1	R@ 2	R@3	Avg R@1
Image Only	18.2	29.6	40.0	9.2	20.2	29.1	9.6	16.2	25.5	6.8	16.0	24.7	11.0
Text Only	12.3	20.2	31.3	8.1	17.7	24.6	8.2	15.3	24.1	7.6	15.4	25.1	9.1
Image+Text	17.6	29.5	40.0	10.6	22.1	31.9	11.8	21.4	29.0	10.3	21.0	31.1	12.6
Combiner-MMC	17.4	29.1	40.5	12.9	22.9	32.4	13.5	23.0	33.3	12.3	22.4	32.1	14.0
FROMAGe-(OPT-6.7B)	**19.2**	**31.1**	**40.5**	12.2	21.7	30.5	13.0	24.5	33.2	12.9	24.7	32.9	14.3
MCL-(OPT-2.7B)	18.2	28.8	38.9	13.9	24.6	34.0	14.6	24.7	34.7	16.8	26.9	36.4	15.8
MCL-(OPT-6.7B)	18.2	29.6	39.1	14.5	24.1	34.0	**14.7**	**27.1**	**36.5**	16.9	28.5	39.3	16.1
MCL-(LLama2-6.7B)	18.5	29.5	38.8	**14.7**	**25.0**	**33.7**	**14.7**	24.9	35.7	**17.2**	**29.8**	**39.9**	**16.3**

Qualitative Results on CIRCO. Figure 5 shows more qualitative results from the CIRCO validation set. The samples from CIRCO are diverse and of high quality. Importantly, it provides multiple ground truth labels for each input, which helps in a more comprehensive analysis. From the figure, we observe that: (1) MCL can effectively handle the multimodal input and retrieve the target image. (2) Some false negative samples are highlighted in blue. This is primarily due to during the label annotation, the authors leverage their proposed SEARLE method to coarsely filter out images from a large gallery, leading to missing some true positives, which can be well retrieved by MCL. These false negatives indicate that MCL has different preferences compared to conventional CLIP inversion-based methods, suggesting MCL's potential to refine existing benchmarks.

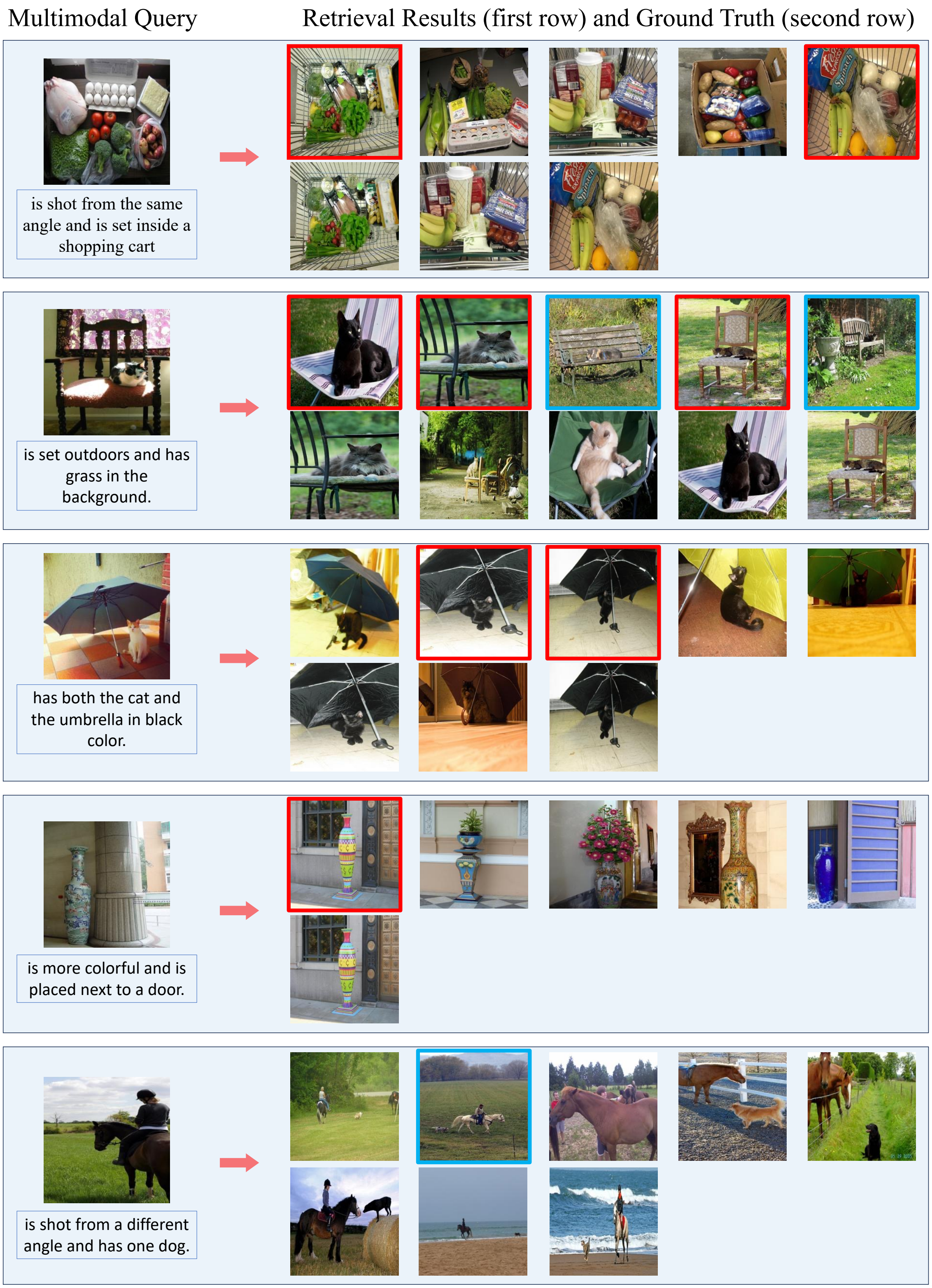


Figure 5. Qualitative results on CIRCO validation set. The first row is the ranked images retrieved by MCL, and the second row is the ground truth images. True positives are marked in red and false negatives are marked in blue.

B. More Qualitative Results and Analysis

B.1. LLM VS. CLIP text encoder

Different from conventional image-text retrieval, logical words like 'no' and 'instead of' occur more frequently in the multimodal contextual image retrieval scenarios. These logical words pose a challenge to CIR models that rely on a CLIP text encoder to process textual queries. It is because the CLIP text encoder struggles in handling words like 'no' (Wang et al., 2023), which are critical to retrieve the correct target images. As shown in Figure 6, the CLIP-based model Combiner struggles to comprehend text inputs containing logical words, leading to incorrect retrieval results. In MCL, we leverage the aligned CLIP image-text space as the retrieval space. The CLIP text encoder is utilized to process the target caption in MMC. In this scenario, the text inputs of CLIP text encoder are more likely in a caption style. The text conditions, which are more likely to contain logical words, are processed in the language model space, where the logical words like 'not' and 'instead of' can be easily understood.

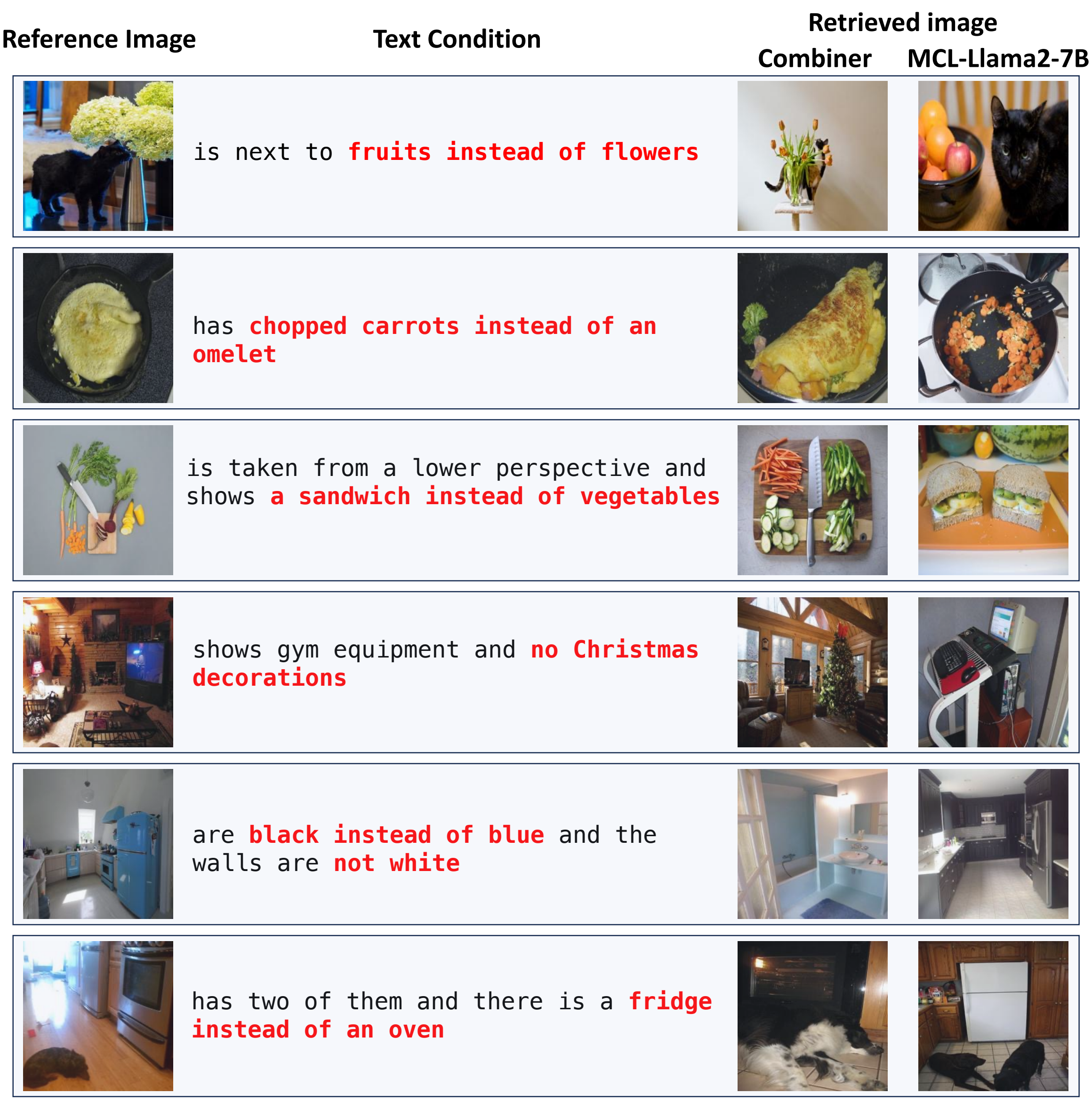


Figure 6. Examples from CIRCO validation set containing logical words. Both Combiner and MCL are trained on proposed MMC dataset.

B.2. Qualitative Results of MCL with Various LMM Backbones

In this section, we provide additional qualitative results to analyze the impact of a stronger language model on MCL. We select some hard samples from CIRCO test set and list the results of MCL with OPT-2.3B, OPT-6.7B and Llama2-7B. As shown in Figure 7, every test sample has a complex text condition, requiring comprehensive multimodal abilities to retrieve the correct image. Overall, benefiting from a more powerful language model, MCL with Llama2-7B shows improved performance in understanding these complex multimodal inputs. For instance, in the first example, Llama2 based model and OPT-6.7B based model identify critical textual cues such as 'is empty' and 'bottles of alcohol next to it' effectively retrieving the correct image, while the OPT-2.7B based models only catch the 'bottles of alcohol' cues.

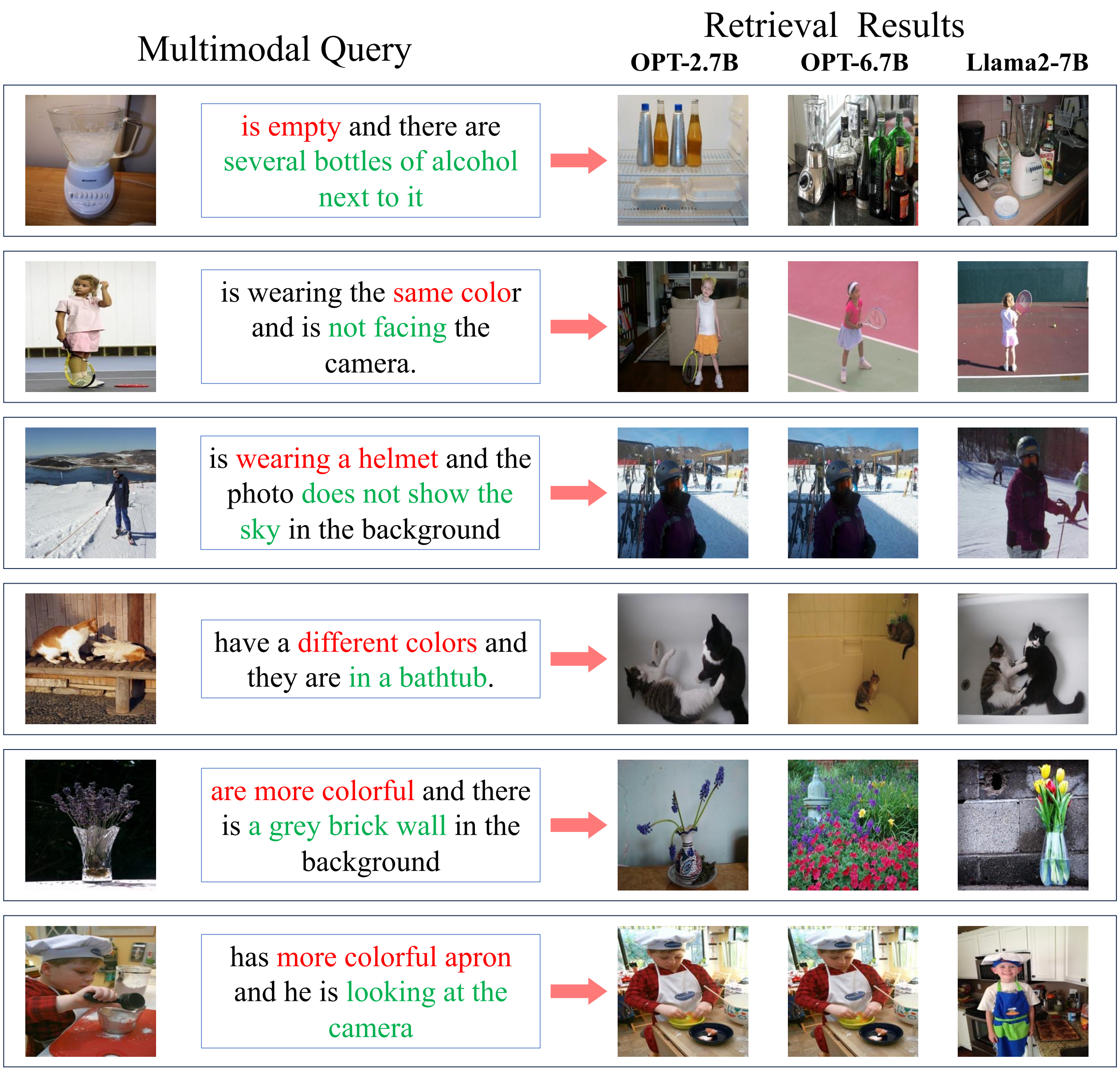


Figure 7. Qualitative comparison of MCL with various LLM backbones on hard samples selected from CIRCO test set. Critical textual cues in each sample are highlighted.

B.3. Qualitative Results on Visual Storytelling

Figure 8 shows the qualitative results of MCL (LLama2) on Visual Storytelling, clearly illustrating the benefits of incorporating more context. For instance, in the final examples, when the input is limited to one caption, the model retrieves a photo that aligns with the caption "cute footprints". However, this photo does not fit the beach theme of the overall visual story. With five captions as input, the model recognizes the beach scenario but displays adult feet, which are inconsistent with the visual context. When the input includes both images and captions, the model accurately retrieves the target image that matches the multimodal context. These findings indicate that MCL can effectively utilize continuous multimodal context, showing great potential for wide-ranging applications in real-world multimodal scenarios.

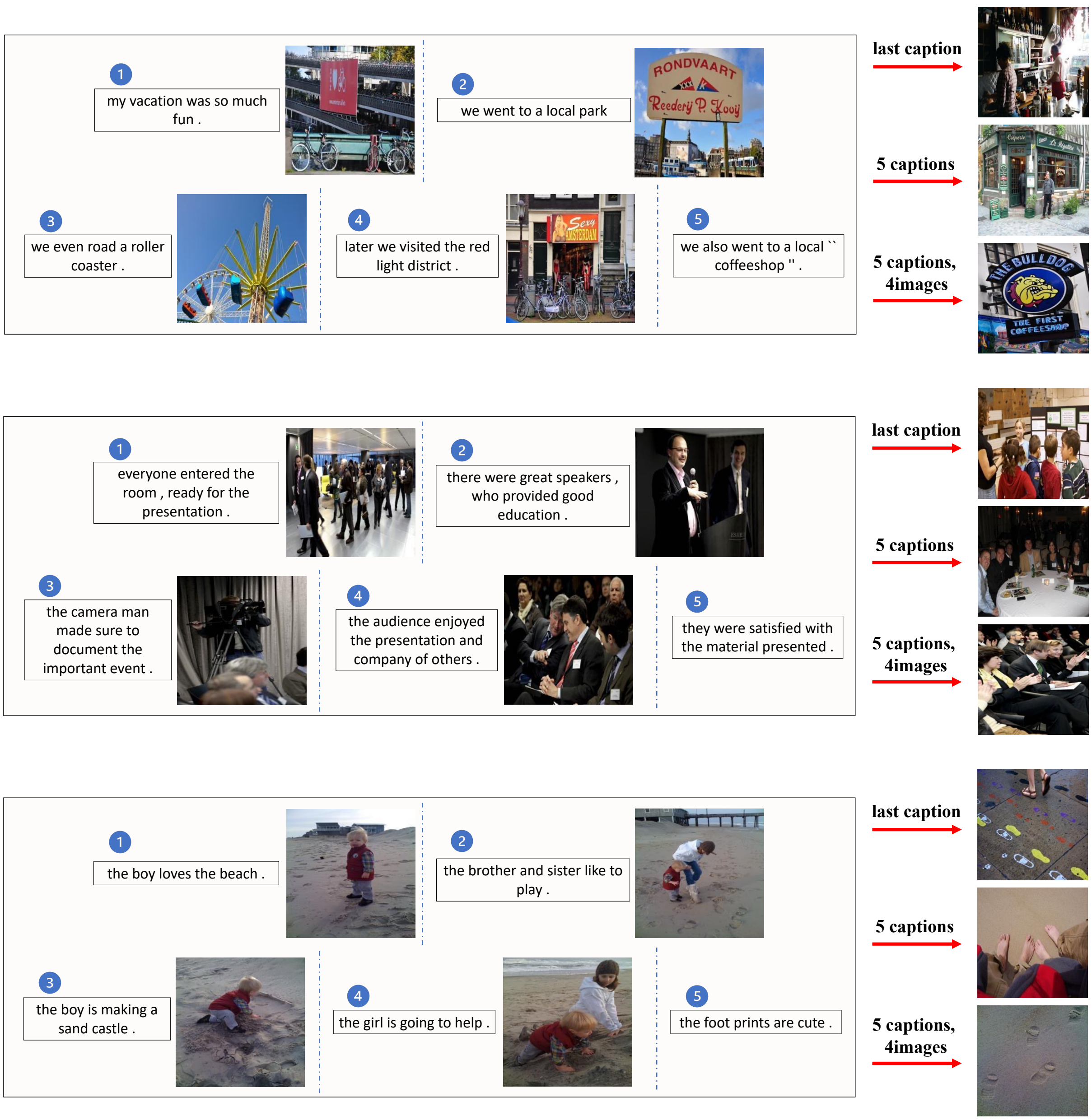


Figure 8. Qualitative Results on Visual StoryTelling.

C. Failure Cases and Limitations

Failure Cases. In Figure 9, we list several primary scenarios where MCL fails. The most significant is the quantities-related scenarios. When the text condition includes specific numerical requirements, such as ‘two cats’, MCL often struggles to retrieve target images with the correct number. Similarly, when the text condition contains words related to quantity, like ‘fewer’ or ‘more’, MCL also fails to accurately identify target images that correctly represent these quantitative relationships. Additionally, MCL struggles with the text condition such as ‘greyscale’ and ‘shot angle’. These text conditions, instead of relating to the image content, are associated with the image’s state or attributes.

Limitations inherited from CLIP model. MCL leverages frozen CLIP model as the base image-text retrieval model. While it benefits from the strong aligned image-text space provided by CLIP, it also inherits CLIP’s inherent limitations. For example, as shown in Figure 9, MCL struggles to retrieve the correct images in scenarios involving queries related to quantities, greyscale and angles. This issue stems from CLIP’s intrinsic weaknesses, such as object counting, as detailed in (Radford et al., 2021). It should be noted that another significant limitation of CLIP is the weak logical understanding ability of CLIP text encoder (Wang et al., 2023). Fortunately, as shown in Figure 6 MCL process the logical relationship in LLM space, thereby mitigating this issue.

Figure 9. Failure cases of MCL.

D. Data Generation

We employ the Llama2/7B-Chat model for our data generation. To enhance the performance of the open-sourced LLM (*i.e.*, Llama2 (Touvron et al., 2023)) in data generation, we utilize the in-context learning techniques. Specifically, we employ a state-of-the-art LLM, GPT-4 , to generate 20 in-context examples. In practice, we find that GPT-4 can effectively perform the data generation task and generate diverse samples. During each sample generation process, we provide Llama2 with a task description and randomly select one in-context example as the task-specific prompt, as shown in Figure 10. This approach ensures diverse and high-quality generated samples. It costs approximately 60 A100 GPU days to generate 2.7 million tuples, using image-caption pairs from CC3M (Sharma et al., 2018) as source pairs. We visualize some samples from MMC as shown in Figure 11. As we can see, Llama2 can generate diverse text conditions based on the source caption. The target caption effectively combines the source caption and the generated text condition.

Figure 10. Our specialized task prompt for Llama2/7B-Chat model.

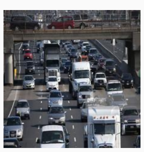
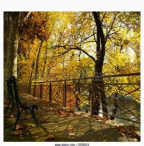
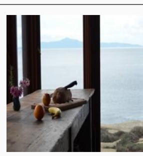
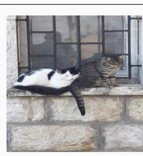
Source image	Source caption	Text Condition	Target Caption
	cartoon smiling moon by the night	showing a sad face	cartoon frowning moon by the night
	amidst the beautiful tea plantations	Add a group of workers picking tea leaves.	Amidst the beautiful tea plantations, a group of workers diligently pick tea leaves.
	it 's about a mile walk around the lake .	Add a bridge crossing the lake.	A mile walk around the lake with a bridge crossing the lake.
	a player works around the cage .	The player should be wearing a different color jersey.	A player works around the cage in green jersey.
	fine line style butterfly tattoo on the right forearm .	remove the butterfly tattoo.	a blank patch of skin on the right forearm.
	traffic makes its way along road .	Add a traffic cone in the middle of the road.	Traffic makes its way along a road with a traffic cone in the middle.
	a park bench under the autumn leaves	Change the season to winter.	a park bench under the winter snow.
	view of the ocean from the kitchen .	a stormy weather.	View of the ocean from the kitchen during a stormy weather.
	cats on the kitchen window	Add a bird in the background.	Cats on the kitchen window with a bird in the background.

Figure 11. Data samples selected from MMC. Note that the source image is not visible to the language model during the text condition and target caption generation.

E. Training details of Combiner baseline

In this paper, we train the Combiner (Baldrati et al., 2022a) using proposed MMC dataset. Given the training data (ref image, text condition, target caption), the Combiner takes reference image and text condition as input. The ref image and text condition are first encoded by CLIP image encoder and CLIP text encoder, respectively. These two clip features are then composed into a single vector through a combiner component (MLPs). A contrastive loss is then used to align the output single vector and the target feature, *i.e.*, the CLIP text feature of the target caption. The Combiner is trained on MMC for 6 epochs with a batchsize 1024. The temperature in contrastive loss is set to 15. The CLIP model is frozen during training. Other training parameters are the same as (Vaze et al., 2023).