Exploring Supervised Demonstration Retrieval For Multimodal In-Context Learning

Anonymous ACL submission

Abstract

Recently, multimodal in-context learning (ICL) has made significant progress, showing impressive performance across various tasks. Existing works demonstrate that demonstration selection have a big influence on the effectiveness of multimodal ICL. However, these methods focus on extracting visual features and textual features from multimodal examples independently and use them for demonstration retrieval. The influence of multimodal embedding methods for ICL demonstration selection is not fully understood. Besides current mulitmodal ICL demonstration retrieval methods are mainly unsupervised, hindering adaptation to specific features of different tasks. To address these challenges, we firstly compare the modality-independent and modality-integrated encoders in representing multimodal examples. Then we introduce MeCO, a supervised training pipeline for multimodal ICL demonstration retriever, cooperating multiple encoders to mitigate their inherent bias and enhance adaptation to specific tasks. Experiments across a wide range of multimodal tasks and MLLMs demonstrate that modality-integrated retrievers show superiority over modality-independent retrievers and our supervised training pipeline significantly improve the performance of multimodal ICL demonstration retrievers which benefit MLLMs on various tasks.

1 Introduction

Recently, multimodal large language models (MLLMs) enable visual understanding and reasoning on complex multimodal tasks (Zhao et al., 2024; Li et al., 2022, 2023b; Liu et al., 2023; Chen et al., 2024b). These models also exhibit in-context learning ability, which has been shown to be largely influenced by the selected in-context learning demonstration examples (Awadalla et al., 2023; Bai et al., 2023; Qin et al., 2024). In LLMs, a series of studies have focused on demonstration retrieval, exploring how to select the optimal examples from the training set for in-context learning (Rubin et al., 2022; Wang et al., 2024a; Li et al., 2023c; Zhang et al., 2023; Qin et al., 2024). In MLLMs, some works study the influence of textual and visual modalities in demonstration selection (Wu et al., 2024a; Jia et al., 2021; Chen et al., 2024a; Luo et al., 2024b).

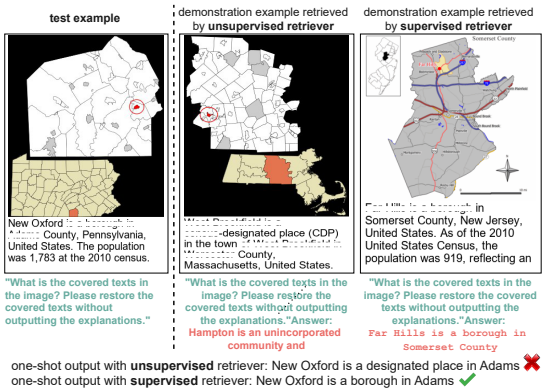

Figure 1: Comparison of ICL demonstration retrieval results with supervised and unsupervised retriever.

However, current demonstration retrieval for MLLMs has two main challenges. First, existing works employ modality-independent retrievers to extract visual and textual features separately for demonstration selection, which hinders the establishment of fused representations for multimodal examples. Recent multimodal embedding methods based on Vision-Language Models (VLMs) provide new possibilities for demonstration retrieval, warranting further investigation. Second, existing mulitmodal in-context learning demonstration retrieval methods are mainly unsupervised, hindering adaption to specific features of different tasks (Qin et al., 2024). Figure 1 demonstrates a case of Visual Caption Restoration (VCR) which involves restoring masked text in images, where the visual elements merely provide supporting context. Given the nature of the task, the retriever should prioritize textual elements within the images over visual fea-

tures. In this example, the demonstration retrieved by unsupervised retriever focus on providing similar visual part, which fails to function as a good one-shot example. In contrast, the demonstration retrieved by supervised retriever contains similar textual parts to the test example, thereby enabling the model to generate the correct response.

To address these two challenges, this paper examines the impact of multimodal encoders on multimodal in-context learning demonstration retrieval and proposes a supervised training pipeline for multimodal demonstration retrievers. For multimodal demonstration embedding, **we compare two methods for embedding multimodal demonstrations:** the modality-independent encoder which treats the text and visual components separately and the modality-integrated encoder which integrates different modalities into a deeply fused representation. **For supervised demonstration retriever training, we propose a Multi-encoder Collaborative Optimization pipeline (MeCO)**. Typical supervised demonstration retriever training pipeline for large language models involves generating a demonstration candidate set, scoring this candidate set, and subsequently training the retriever through contrastive learning. MeCO leverages the varied recall results of different encoders to provide a high-quality set of positive candidates. Our pipeline mitigates the potential bias on candidates selection brought by single-way candidate recall, enabling the encoders to learn from each other.

We conduct experiments on a wide range of multimodal tasks and MLLMs. Experiments show that the modality-integrated retrievers trained with the MeCO pipeline significantly improve the performance of MLLMs in multiple tasks. We summarize the contribution of this as follows.

(1) We comprehensively evaluate the performance of modality-independent retriever versus modality-integrated retriever in demonstration retrieval for multimodal in-context learning. Modality-integrated retrievers outperform modality-integreated retriever, especially on the challenging tasks. Moreover, modality-independent retriever tends to learn spurious features during supervised training while modality-integrated retrievers are more robust.

(2) We propose Multi-encoder Collaborative Optimization (MeCO), a supervised in-context learning demonstration retrieval method for MLLMs, which cooperates multiple encoders to mitigate their inherent bias and enhance adaptation to specific tasks.

2 Related Works

2.1 Multimodal In-Context Learning

In-context learning, a crucial capability of LLMs, is also considered as important for MLLM. A series of MLLMs successfully inherit in-context learning capabilities by employing various techniques during pre-training and fine-tuning stage (Huang et al., 2023; Laurençon et al., 2024; Bai et al., 2023). These techniques include constructing interleaved image-text training data and instruction tuning (Li et al., 2023a), multi-turn curriculum-based learning methodology with effective data mixes (Doveh et al., 2024), and compacting the latent space of visual prompts (Gao et al., 2024). Recently Qin et al. (2024) propose a general analysis of the underlying factors affecting the effectiveness of multimodal in-context learning, including multimodal demonstration retrieval, intra-demonstration ordering, and the introductory instructions in prompts. Our work differs from Qin et al. (2024) by focusing on harnessing demonstration retrieval to boost multimodal in-context learning and propose a supervised training pipeline .

2.2 Demonstration Retrieval

Existing studies have witnessed the huge impact of in-context examples selection on LLMs' performance (Liu et al., 2022; Luo et al., 2024a; Agrawal et al., 2023). A line of studies then focused on finding good in-context examples by representing examples with dense encoder and choosing the semantically similar ones (Rubin et al., 2022; Li et al., 2023c; Liu et al., 2022). Rubin et al. (2022) use unsupervised encoder to get candidates and then generate training data by scoring them to train a supervised demonstration retriever. A series of works follows Rubin et al. (2022), such as proposing a unified demonstration retriever and expanding this procedure to visual in-context learning (Li et al., 2023c; Liu et al., 2022; Wang et al., 2024a; Luo et al., 2024b). In this paper, we adapt this training pipeline in the realm of multimodal tasks to boost the performance of MLLMs.

3 Method

3.1 Problem Definition

The goal of multimodal in-context learning demonstration retrieval is to retrieve the most suitable examples from the training set D to construct the in-context learning prompt $\mathcal{P} = \{(x_i, y_i)\}_{i=1}^{m} \subseteq D$,

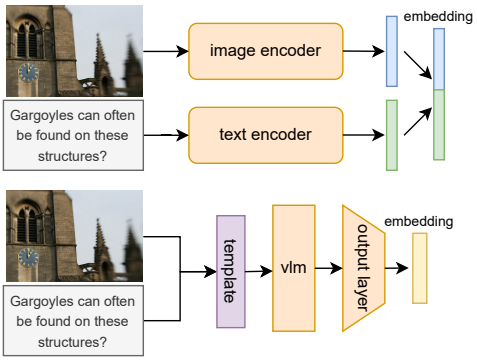

Figure 2: Multimodal demonstration embedding methods.

for a given test example x_q. We formulate this problem in the context of multimodal task as:

$$(x^I, x^T) = \arg \max_{(x_n^I, x_n^T) \in D} f_\theta((x_n^I, x_n^T), (x_q^I, x_q^T)). \quad (1)$$

where x^I and x^T represent respectively the visual part and textual part of the examples and f_θ is the retrieval function parameterized by θ. In this work, we implement f_θ as the cosine similarity of dense embeddings of the query and training examples. The two critical steps of a demonstration retrieval is (1) representing the multimodal examples as dense embeddings, incorporating information from both modalities, and (2) fine-tuning the representation space according to different tasks to explicitly optimize it for multimodal in-context learning.

3.2 Mulitmodal Demonstration Embedding

In this section, we compare two methods to represent multimodal examples considering the two modalities as showed in Figure 2.

The modality-independent embedding calculates the image embedding and text embedding with a CLIP-like encoder and concatenate the embedding together to get the multimodal example representation. In this case, the embeddings of the two modalities are independent and final example similarity is equivalent to the sum of visual similarity and textual similarity as showed in Equation 2.

$$\begin{aligned} \text{Sim}(E_i, E_j) &= \text{Sim}([E_i^T, E_i^I], [E_j^T, E_j^I]) \\ &= \text{Sim}(E_i^T, E_j^T) + \text{Sim}(E_i^I, E_j^I) \end{aligned} \quad (2)$$

The modality-integrated embedding aims to represent the image, text, and the task instruction as an integrated embedding. As illustrated in Figure 2, we employ an embedding model based on a vision-language model. The entire example is structured into a template and then embedded using the last token of the vision-language model, which is followed by an output layer. The advantage of this approach is that the image features and text features are deeply integrated within the transformer architecture, enabling better capture of cross-modal relationships. Furthermore, this embedding model can process input with various combinations of images and texts, thereby aligning the task description to establish a task-specific embedding.

3.3 MeCO Pipeline

In this section, we present our MeCO pipeline to jointly train multimodal in-context learning demonstration retrievers. We want to explicitly optimize the demonstration retrievers so that the selected examples based on Equation 1 can maximize few-shot performance of multimodal large language models. In Section 3.3.1, we first introduce how to obtain positive-negative example pairs for each query as training data through the cooperation of multiple encoders. Then, in Section 3.3.2, we describe how to train the retrievers using this data.

3.3.1 Generating Training Data

Our evaluation of the two demonstration embedding models indicates that both can enhance the few-shot performance of multimodal large language models. Notably, we observed a small overlap in their retrieval results, suggesting that different multimodal embedding models can identify distinct, characteristic demonstrations. This finding motivates us to conduct a multi-faceted candidates recall process by integrating different demonstration embedding models to retrieve a diverse candidate set. This strategy facilitates mutual learning among encoders and mitigates the inherent biases in their demonstration selection process.

Specifically, for a given query from the training set, we first utilize each unsupervised retriever to recall a candidate set of demonstrations with the top-k cosine-similarity scores relative to the query. Then we combine these results and augment them with randomly sampled examples from the training set to increase diversity.

$$s(\mathcal{P}) = \log p(y_q | \mathcal{P}, x_q). \quad (3)$$

$$s(\mathcal{P}) = \text{metric}(y_q, G_\Theta(\mathcal{P}, x_q)). \quad (4)$$

Subsequently, we evaluate all the demonstrations in the candidate set using Equation 3 or 4 . x_q and y_q in Equation 3 and 4 denote respectively the input and ground truth of the query, $\mathcal{P}$ denotes the

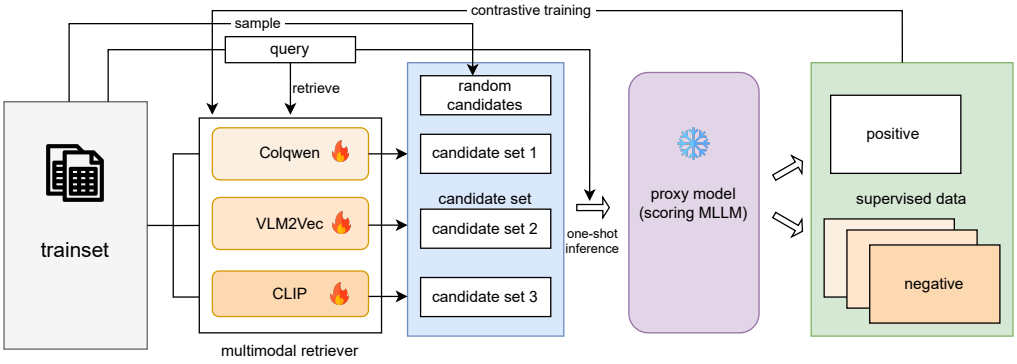

Figure 3: Overview of MeCO pipeline. ColQwen and VLM2Vec are modality-integrated retriever and CLIP is modality-independent retriever.

one-shot example and G_Θ represents the MLLM. The evaluation is performed by inputting the examples into a proxy model, a smaller multimodal large language model. Equation 3 computes the score as the log-likelihood of generating the ground-truth when using the chosen example as a prompt, while Equation 4 directly decodes the answer and employs the resulting task-specific metric as the score. Empirically, we find that score calculation using Equation 4 yields superior performance. This superiority can be attributed to its better alignment with the evaluation criteria of each task, particularly for text generation tasks where exact matching with the ground truth is not essential.

Finally, we select the top-5 scoring examples as public positive examples. For each retriever we identify its hard negative examples by intersecting the set of bottom-5 scoring examples with the examples retrieved by itself. This approach, which incorporates all unsupervised retrievers, allows us to construct a high-quality set of public positive examples and several personalized sets of hard negative examples for each demonstration retriever.

3.3.2 Contrastive Learning of Retrievers

Now we have a supervised trainset for demonstration retriever D_{sup} built from the original trainset D. For each example q_i in the trainset D, we have a set a positive examples $\xi_{pos}(i)$ and a set of hard negative examples $\xi_{neg}(i)$).

$$D_{sup} = \{(q_i, \xi_{pos}(i), \xi_{neg}(i)) | \forall q_i \in D\}.$$

We train the retriever with typical contrastive learning objective. For each training instance in a mini-batch of size B, we sample one positive example d_i^+ from the positive set $\xi_{pos}(i)$ and one hard negative d_i^- from the negative set $\xi_{neg}(i)$, which consists the negative examples with the other $B-1$ positive examples in the same min-batch. The final contrastive loss is computed as:

$$\ell = -\log \frac{\exp(s(q_i, d_i^+))}{\exp(s(q_i, d_i^-)) + \sum_{j=1, j\neq i}^{B} \exp(s(q_i, d_j^+))}. \quad (5)$$

$$s(q, d) = E_\theta(q)^T E_\theta(d)/\tau. \quad (6)$$

where we have the similarity of a query and a demonstration is the inner-product of their embeddings scaled by a temperature τ.

4 Experimental Results

In this section we first comprehensively compare the performance of different unsupervised multimodal retrievers. Then we evaluate the performance of supervised demonstration retrievers trained with MeCO pipeline.

Dataset We conduct experiments on a series of multimodal tasks including Visual Question Answering (VQA), Visual Captioning, and Visual Caption Restoration (VCR).

VQA We include three traditional VQA datasets, Vizwiz (Gurari et al., 2018), OK-VQA (Marino et al., 2019), and VQAv2 (Goyal et al., 2017), using BertScore as the metric. We also include Hateful Memes (Kiela et al., 2020), which focuses on detecting hateful speech in multimodal memes. We use AUC-ROC as the metric.

Visual Caption We include Flickr30k (Young et al., 2014) and use CIDEr as the metric.

Visual Caption Restoration VCR challenges models to restore partially obscured text within images, leveraging pixel-level hints and contextual cues from the image (Zhang et al., 2024). It includes two languages, English (VCR-en) and Chinese (VCR-zh). VCR-zh poses greater challenges due to the complexity of Chinese character. For

		VizWiz	VQAv2	OK-VQA	VCR-en	VCR-zh	Flickr30k	Hateful	avg
Qwen2-VL -7B	Zero-shot	62.4	89.5	67.6	76.5	60.3	79.5	66.0	71.7
	Random	62.8	90.5	70.8	83.2	63.7	**82.6**	66.9	74.4
	CLIP-ViT	66.4	91.0	70.9	84.2	69.9	81.2	**71.3**	76.4
	VLM2Vec	**66.8**	91.0	**71.2**	84.9	–	80.7	70.2	77.5
	ColQwen	66.6	**91.8**	71.0	**85.0**	**76.9**	80.8	70.9	77.6
DeepSeek -VL2-7B	Zero-shot	52.8	75.6	70.9	41.3	0.0	54.5	62.1	51.0
	Random	54.9	73.8	73.6	40.9	15.3	57.3	62.2	54.0
	CLIP-ViT	**60.6**	78.4	74.3	48.5	30.2	60.4	64.7	59.6
	VLM2Vec	59.9	**82.4**	74.5	47.3	–	61.4	64.9	65.1
	ColQwen	59.5	81.1	**75.0**	**50.9**	**50.4**	**61.8**	**65.5**	63.5
Claude3.5 -Sonnet	Zero-shot	30.0	58.8	31.7	61.6	0.04	25.8	73.9	40.3
	Random	40.8	79.9	50.4	68.3	0.06	**35.8**	73.7	49.9
	CLIP-ViT	46.2	80.2	51.9	71.9	27.2	35.1	76.2	55.5
	VLM2Vec	**46.6**	81.1	**53.1**	74.2	–	34.7	**76.9**	61.1
	ColQwen	45.3	**81.2**	53.0	**74.6**	**48.4**	34.0	75.6	58.9
GPT4o -0513	Zero-shot	51.0	67.6	42.4	78.2	11.1	47.7	75.6	53.4
	Random	55.5	85.5	59.8	83.8	17.5	**57.4**	75.6	62.2
	CLIP-ViT	55.7	86.1	61.0	84.1	35.3	55.4	77.3	65.0
	VLM2Vec	**56.4**	**86.8**	**63.8**	**84.8**	–	56.0	**77.6**	70.9
	ColQwen	56.1	**86.8**	62.3	**84.8**	**50.4**	55.0	77.3	67.5

Table 1: Zero-shot and in-context learning performance of MLLMs with different **unsupervised demonstration retrievers**. The VLM2Vec model does not support Chinese, so we mask its performance on VCR-zh.

this visual cloze task, we use exact match accuracy as the metric.

Retrievers and Models We conduct experiments with one modality-independent retriever based on CLIP-ViT-Large (Radford et al., 2021), and two modality-integrated retrievers VLM2Vec (Jiang et al., 2024) and ColQwen (Faysse et al., 2024). VLM2Vec process any combination of images, text and task description to generate an embedding using the final token's hidden state from a vision-language model. ColQwen leverages contextualized embeddings from Qwen2-VL along with a late interaction matching mechanism for visual document retrieval. We assess the performance of these retrievers across various multimodal large language models, including both closed-source models (GPT-4o-0513 and Claude3.5-Sonnet) and open-source models (Qwen2-VL-7B (Wang et al., 2024b), and DeepSeek-VL2 (Wu et al., 2024b)).

4.1 Unsupervised Retrieval Results

Main results Table 1 shows that demonstrations retrieved by all three unsupervised retrievers significantly enhance the model's few-shot performance compared to randomly selected demonstrations. Meanwhile **modality-integrated encoders outperform modality-independent encoders in demonstration retrieval capability**. Specifically, Qwen2-VL achieves an average few-shot performance of 77.6 across all tasks using ColQwen-retrieved demonstrations, while the performance drops to 76.4 when using CLIP-ViT for retrieval. Notably, the performance differential between these two retrieval approaches exhibits task-specific and model-specific variations.

Analysis for different models Notably, the two closed-source models demonstrate superior performance on challenging tasks such as VCR and Hateful Memes, which require sophisticated visual-linguistic understanding and reasoning. However, they show relatively lower performance compared to open-source models on conventional tasks like VQA and Visual Caption. This performance discrepancy can be attributed to their tendency to generate more elaborate and nuanced responses, which ironically becomes disadvantageous when handling simpler tasks that require simple and straightforward answers. Nevertheless, this observation does not affect our primary analysis, which focuses on comparing the effectiveness of different demonstration retrieval strategies rather than conducting cross-model evaluations.

Analysis for different tasks We find that the impact of demonstration selection varies across different tasks. **For VQA tasks, demonstration retrieval shows relatively small impact on model performance**, primarily because current VQA datasets mainly evaluate basic image understanding capabilities. In this context, demonstrations

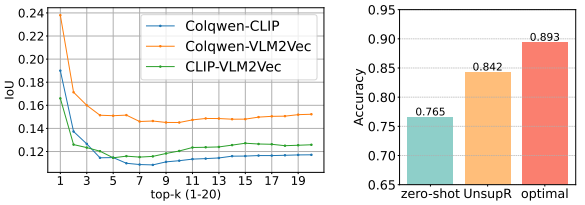

Figure 4: Left: The IoU of the retrieval results of different demonstration encoders. Right: Exact match accuracy on vcr-en-hard in zero-shot and in-context learning settings.

	Viz.	vcr-e.	hate.	Flickr.
Qwen2-VL-7B				
Random	62.8	83.2	66.9	82.6
CLIP-ViT	66.7$_{\uparrow 0.3}$	84.6$_{\uparrow 0.4}$	70.4$_{\downarrow 0.9}$	82.6$_{\uparrow 1.4}$
VLM2Vec	**67.4**$_{\uparrow 0.6}$	**86.5**$_{\uparrow 1.6}$	71.2$_{\uparrow 1.1}$	83.3$_{\uparrow 2.6}$
ColQwen	67.3$_{\uparrow 1.3}$	86.4$_{\uparrow 1.4}$	**71.8**$_{\uparrow 0.9}$	**83.5**$_{\uparrow 2.7}$
DeepSeek-VL2				
Random	55.5	40.9	62.1	57.3
CLIP-ViT	62.1$_{\uparrow 1.5}$	51.9$_{\uparrow 3.4}$	64.1$_{\downarrow 0.6}$	62.6$_{\uparrow 2.2}$
VLM2Vec	61.9$_{\uparrow 2.0}$	52.9$_{\uparrow 5.6}$	66.3$_{\uparrow 1.4}$	62.0$_{\uparrow 0.4}$
ColQwen	**62.2**$_{\uparrow 2.7}$	**55.6**$_{\uparrow 4.7}$	**66.9**$_{\uparrow 1.4}$	**64.0**$_{\uparrow 2.2}$
GPT4o-0513				
Random	55.5	83.8	75.6	57.4
CLIP-ViT	56.1$_{\uparrow 0.4}$	84.6$_{\uparrow 0.5}$	78.0$_{\uparrow 0.3}$	55.6$_{\uparrow 0.2}$
VLM2Vec	56.7$_{\uparrow 0.3}$	85.6$_{\uparrow 0.8}$	**78.3**$_{\uparrow 0.7}$	**55.8**$_{\downarrow 0.2}$
ColQwen	**56.8**$_{\uparrow 0.7}$	**85.9**$_{\uparrow 1.1}$	77.2$_{\downarrow 0.1}$	55.4$_{\uparrow 0.4}$
Claude3.5-Sonnet				
Random	45.8	68.3	73.7	35.8
CLIP-ViT	46.7$_{\uparrow 0.5}$	73.3$_{\uparrow 1.4}$	76.0$_{\downarrow 0.2}$	35.4$_{\uparrow 0.3}$
VLM2Vec	46.3$_{\downarrow 0.3}$	73.7$_{\downarrow 0.5}$	**77.0**$_{\uparrow 0.1}$	**36.4**$_{\uparrow 1.7}$
ColQwen	**47.4**$_{\uparrow 1.9}$	**75.4**$_{\uparrow 0.8}$	75.8$_{\uparrow 0.2}$	35.7$_{\uparrow 1.7}$

Table 2: Few-shot performance of multimodal large language models with supervised demonstration retrievers. **The number ↑ indicates the performance improvement from MeCO pipeline.**

primarily serve to regularize responses into concise phrases, particularly for closed-source models, an effect achievable even with random demonstrations. In contrast, **demonstration selection significantly impacts model performance on Visual Caption Restoration (VCR) and Hateful Memes tasks.** These tasks pose substantial challenges, requiring models to process pixel-level details and perform complex visual-linguistic reasoning. For such challenging tasks, modality-integrated retrievers consistently outperform their modality-independent counterparts in demonstration selection, leading to notable performance improvements. For **image caption tasks**, our experiments reveal that **retrieved demonstrations prove less effective than randomly selected samples** for in-context learning. This observation aligns with Yang et al. (2023), who found that visually similar images may not provide optimal demonstrations and could potentially interfere with the generation of accurate descriptions for the target image.

The above analysis demonstrates that the criteria for effective demonstrations vary across different tasks, emphasizing the importance of task-specific retriever optimization.

Comparison of retrieval outcomes To compare the retrieval outcomes from the three multimodal retrievers, we computed the Intersection over Union (IoU) of the top-k candidates recalled by each retriever on VCR-en. As depicted in Figure 4, the three retrieval methods demonstrate limited overlap in their retrieval results, and the IoU drops to less than 0.15 when top-k is above 5. **This indicates that each retriever identifies distinctive demonstrations for multimodal in-context learning.**

To assess the upper bound of retrieval effectiveness, we manually identify the most beneficial demonstration from the top-20 candidates retrieved by each retriever for each test example. Figure 4 compares the performance on VCR-en under zero-shot and one-shot settings, contrasting results between unsupervised retrieval and these manually selected optimal demonstrations. While multimodal large language models show substantial improvements with demonstrations from unsupervised retrievers, there remains **a considerable performance gap between unsupervised retrieval and the achievable upper bound.** These findings underscore the potential benefits of supervised retriever fine-tuning.

4.2 Supervised Retrieval Results (MeCO)

Setup We evaluate our supervised retrievers on four tasks, Vizwiz, VCR, HatefulMemes and Flickr30k. In our training pipeline, we employed Qwen2-VL-2B as the proxy for Qwen2-VL-7B. DeepSeek-VL served as its own proxy model thanks to its efficient inference capabilities. For the two closed source model, we opted for Qwen2-VL-2B as the proxy model. More experimental details are shown in the Appendix A.2.

Results Table 2 demonstrates that all three retrievers can benefit from our MeCO pipeline and retrieve better prompts compared to unsupervised ones. For image caption task, supervised retrievers can lead to a performance boost of average 2.3 for

	Viz.	vcr-e.	hate.	Flickr.
proxy: DeepSeek-VL2 (itself)				
CLIP-ViT	$62.1_{\uparrow 1.5}$	$51.9_{\uparrow 3.4}$	$64.1_{\downarrow 0.6}$	$62.6_{\uparrow 2.2}$
VLM2Vec	$61.9_{\uparrow 2.0}$	$52.9_{\uparrow 5.6}$	$66.3_{\uparrow 1.4}$	$62.0_{\uparrow 0.4}$
ColQwen	$\mathbf{62.2}_{\uparrow 2.7}$	$\mathbf{55.6}_{\uparrow 4.7}$	$\mathbf{66.9}_{\uparrow 1.4}$	$\mathbf{64.0}_{\uparrow 2.2}$
proxy: Qwen2-VL-2B				
CLIP-ViT	$61.9_{\uparrow 1.3}$	$45.4_{\downarrow 3.1}$	$64.8_{\uparrow 0.1}$	$60.5_{\uparrow 0.1}$
VLM2Vec	$60.4_{\uparrow 0.5}$	$49.4_{\uparrow 2.1}$	$65.1_{\uparrow 0.2}$	$60.7_{\downarrow 0.7}$
ColQwen	$61.6_{\uparrow 2.1}$	$51.5_{\uparrow 0.6}$	$65.3_{\downarrow 0.2}$	$62.4_{\uparrow 0.6}$

Table 3: The impact of proxy model selection to MeCO pipeline. We use DeepSeek-VL2 as the multimodal large language model.

	Viz.	vcr-e.	hate.	Flickr.
MeCO: multi-way candidate recall				
CLIP-ViT	$66.7_{\uparrow 0.3}$	$84.6_{\uparrow 0.6}$	$70.4_{\downarrow 0.9}$	$82.6_{\uparrow 1.4}$
VLM2Vec	$\mathbf{67.4}_{\uparrow 0.6}$	$\mathbf{86.5}_{\uparrow 1.6}$	$71.2_{\uparrow 1.1}$	$83.3_{\uparrow 2.6}$
ColQwen	$67.3_{\uparrow 1.3}$	$86.4_{\uparrow 1.4}$	$\mathbf{71.8}_{\uparrow 0.9}$	$\mathbf{83.5}_{\uparrow 2.7}$
traditional EPR: single-way candidate recall				
CLIP-ViT	$65.9_{\downarrow 0.5}$	$84.6_{\uparrow 0.6}$	$70.5_{\downarrow 0.8}$	$81.8_{\uparrow 0.6}$
VLM2Vec	$67.0_{\uparrow 0.2}$	$85.5_{\uparrow 0.6}$	$70.8_{\uparrow 0.7}$	$81.6_{\uparrow 0.9}$
ColQwen	$67.1_{\uparrow 1.1}$	$85.7_{\uparrow 0.7}$	$71.2_{\uparrow 0.3}$	$81.9_{\uparrow 0.7}$

Table 4: The influence of using MeCO. We use Qwen2-VL as the multimodal large language model.

Qwen2-VL compared with unsupervised retrievers and outperforms random demonstrations.

Results also indicate that supervised training demonstrates greater effectiveness for modality-integrated retrievers than for their modality-independent counterpart. Specifically, when tested with Deepseek2-VL, the modality-integrated retrievers ColQwen and VLM2Vec show substantial improvements after MeCO training, achieving average performance gains of 2.7 and 2.4 points respectively across all four tasks. Notably the modality-independent CLIP-ViT exhibits a smaller improvement of 1.6 points. Similar patterns are observed in experiments with Qwen2-VL. This stronger benefit from supervised training stems from modality-integrated retrievers' capacity to generate unified representations that deeply fuse information from images, texts, and task descriptions. **This comprehensive multimodal fusion enables better understanding of task-specific features** during training, thus facilitating more effective fine-tuning.

Influence of Proxy Model We find that the two closed-source model using Qwen2-VL-2B as proxy model can also benefit from the MeCO training pipeline, but performance gains are smaller than Qwen2-VL and DeepSeek-VL2. To study the influence of proxy model selection to the MeCO training pipeline, we employ respectively Qwen2-VL-2B and DeepSeek-VL2 as the proxy model for DeepSeek-VL2 and compare the final few-shot learning performance. Table 3 reveals **the importance of using a proxy model with the same architecture in MeCO pipeline**. When optimizing retrievers for DeepSeek-VL2, ColQwen achieves an average improvement of 2.8 points when trained with the same proxy model, while this improvement drops significantly to 0.8 points when using a different proxy model, Qwen2-VL-2B. This substantial performance gap arises from the architectural differences between the models, as DeepSeek-VL2 employs MoE architecture while Qwen2-VL does not.

Multi-encoder Collaboration We conducted a comparative analysis between retrievers trained using our MeCO pipeline and those trained with the traditional Efficient Prompt Retriever (EPR) pipeline (Rubin et al., 2022). In the EPR approach, each retriever independently recalls its candidate set, which is then processed by the proxy model to generate supervised training data. As shown in Table 4, MeCO consistently improves the performance across all three retrievers. Notably, on the Flickr30k dataset, MeCO-trained retrievers outperform random demonstration selection, while EPR-trained retrievers fail to surpass this baseline.

5 Further Analysis

5.1 Number of In-Context Examples

We investigate the relationship between the number of in-context learning examples and the performance of multimodal large language models. Specifically, we examine Qwen2-VL's performance on two distinct tasks: VCR-en and Flickr30k, utilizing ColQwen as the multimodal retriever. The results, as illustrated in Figure 6, reveal that the impact of in-context learning examples varies significantly across different tasks. For the VCR task, we observe that one-shot yields already substantial performance improvements, while additional examples showing small returns. In contrast, the Visual Caption task demonstrates a consistent positive correlation between performance and the number of in-context learning examples, showing steady improvements as more examples are added. Notably, our supervised retriever trained with our MeCO pipeline effectively enhances few-shot performance

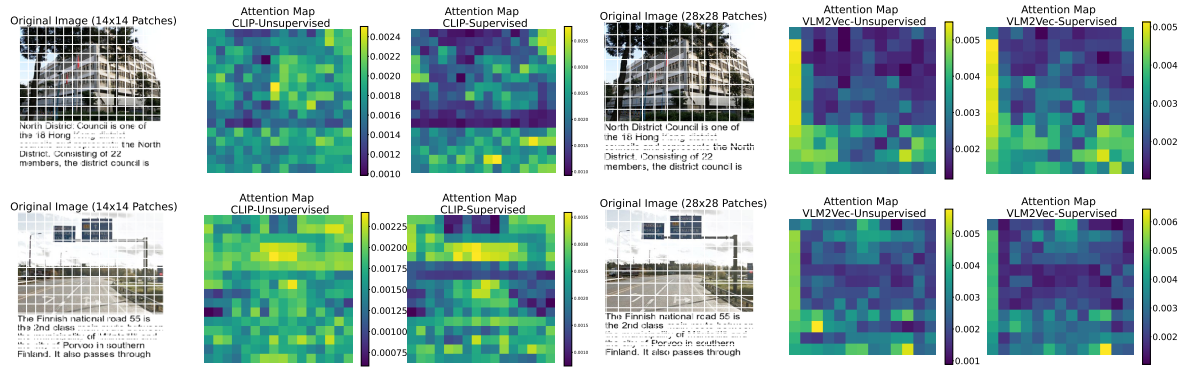

Figure 5: The attention map of image tokens to the pooling tokens. We use the attention scores in the first head of the first layer.

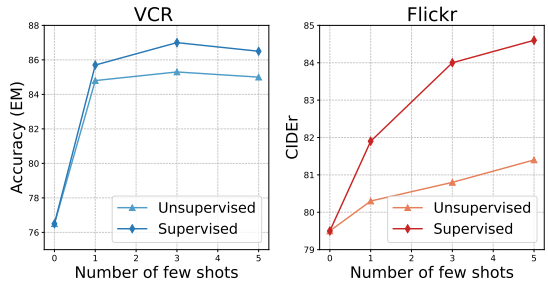

Figure 6: Impact of in-context learning example number with unsupervised retriever and supervised retriever.

across both tasks, maintaining its efficacy across varying numbers of in-context examples.

5.2 How Supervised Training Affects Multimodal Embedding

To better understand the differences between modality-independent and modality-integrated retrievers, as well as their changes after MeCO training, we present attention maps showing how image tokens attend to the pooling token in CLIP-ViT and VLM2Vec. ColQwen is excluded from this analysis as it employs a late interaction mechanism to compute the final embedding, rather than utilizing a pooling token. The example images are selected from the VCR-en dataset, where identifying demonstrations with similar visual characters (shown in the lower portion of the image) is crucial, while the background picture in the upper portion serves as contextual reference.

As shown in Figure 8, the unsupervised CLIP-ViT retriever fails to allocate sufficient attention to visual characters in the image. After MeCO training this region receives notably increased attention, which explains the model's enhanced ability to identify relevant demonstrations for the VCR task. The second example demonstrates a failure case of the CLIP retriever trained with MeCO, where CLIP-ViT predominantly focuses on the white regions of the image instead of the relevant visual characters. This case along with more examples presented in the Appendix B.2 , suggests that the CLIP-ViT retriever trained on VCR-en learns spurious correlations by incorrectly attributing importance to white areas (which happen to coincide with character regions). This observation highlights a fundamental limitation of modality-independent retriever: without proper cross-modal fusion guided by natural language task descriptions, they are more susceptible to learning spurious features during supervised training in the MeCO pipeline.

In contrast, VLM2Vec demonstrates stronger attention to character-containing regions even before MeCO training, attributed to its unified representation that incorporates task descriptions from the text input. The attention pattern remains relatively consistent after MeCO training.

6 Conclusion

This study investigates demonstration retrieval strategy for multimodal in-context learning. We find that the modality-integrated retriever has superior performance to the modality-independent retriever. We also propose MeCO, the first multimodal in-context learning demonstration retriever training pipeline, which make three retrievers to jointly recall candidate set to optimize for different tasks. Experiments show that the MeCO pipeline significantly improve the in-context learning example retrieval abilities of multimodal demonstration retrievers. We also visualize the differences between the modality-independent and modality-integrated retrievers.

7 Limitations

The proposed MeCO training pipeline in this paper relies on proxy models for scoring candidate sets, which is resource-intensive. This is especially true for closed-source models like GPT-4o, where utilizing the model itself as a proxy to label training samples incurs substantial costs. This issue is a common challenge for existing in-context learning demonstration retrieval training methods. Future research could explore in-context learning demonstration retrieval approaches that do not depend on proxy model supervision, thereby reducing training costs.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8857–8873. Association for Computational Linguistics.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *CoRR*, abs/2308.01390.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Shuo Chen, Zhen Han, Bailan He, Jianzhe Liu, Mark Buckley, Yao Qin, Philip Torr, Volker Tresp, and Jindong Gu. 2024a. Can multimodal large language models truly perform multimodal in-context learning? *Preprint*, arXiv:2311.18021.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Preprint*, arXiv:2312.14238.

Sivan Doveh, Shaked Perek, Muhammad Jehanzeb Mirza, Amit Alfassy, Assaf Arbelle, Shimon Ullman, and Leonid Karlinsky. 2024. Towards multimodal in-context learning for vision & language models. *CoRR*, abs/2403.12736.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *CoRR*, abs/2407.01449.

Jun Gao, Qian Qiao, Ziqiang Cao, Zili Wang, and Wenjie Li. 2024. Aim: Let any multi-modal large language models embrace efficient in-context learning. *arXiv preprint arXiv:2406.07588*.

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 *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3608–3617. Computer Vision Foundation / IEEE Computer Society.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Johan Bertil Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2024. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *CoRR*, abs/2410.05160.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *CoRR*, abs/2405.02246.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. MIMIC-IT: multi-modal in-context instruction tuning. *CoRR*, abs/2306.05425.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023c. Unified demonstration retriever for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4644–4668. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *CoRR*, abs/2304.08485.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, DeeLIO@ACL 2022, Dublin, Ireland and Online, May 27, 2022*, pages 100–114. Association for Computational Linguistics.

Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. 2024a. In-context learning with retrieved demonstrations for language models: A survey. *CoRR*, abs/2401.11624.

Yang Luo, Zangwei Zheng, Zirui Zhu, and Yang You. 2024b. How does the textual information affect the retrieval of multimodal in-context learning? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5321–5335. Association for Computational Linguistics.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE.

Libo Qin, Qiguang Chen, Hao Fei, Zhi Chen, Min Li, and Wanxiang Che. 2024. What factors affect multi-modal in-context learning? an in-depth exploration. *CoRR*, abs/2410.20482.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2655–2671. Association for Computational Linguistics.

Huazheng Wang, Jinming Wu, Haifeng Sun, Zixuan Xia, Daixuan Cheng, Jingyu Wang, Qi Qi, and Jianxin Liao. 2024a. MDR: model-specific demonstration retrieval at inference time for in-context learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4189–4204. Association for Computational Linguistics.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A. Rossi, Ruiyi Zhang, Subrata Mitra, Dimitris N. Metaxas, Lina Yao, Jingbo Shang, and Julian J. McAuley. 2024a. Visual prompting in multimodal large language models: A survey. *CoRR*, abs/2409.15310.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024b. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. 2023. Exploring diverse in-context configurations for image captioning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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

Tianyu Zhang, Suyuchen Wang, Lu Li, Ge Zhang, Perouz Taslakian, Sai Rajeswar, Jie Fu, Bang Liu, and Yoshua Bengio. 2024. VCR: visual caption restoration. *CoRR*, abs/2406.06462.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. 2023. What makes good examples for visual in-context learning? In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2024. MMICL: empowering vision-language model with multi-modal in-context learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

A Experimental Details

A.1 Dataset and Metrics

To control computational costs, we sliced larger datasets into subsets to match our computational resource levels. For VQAv2, we take the first 2500 examples in the original test set for evaluation. For VCR-en we use the official test subset of 500 examples for evaluation and take the first 100k examples in the train set for in-context learning demonstration retriever training. For the other datasets used in this paper, we use the original test set for evaluation and train set for the training of the in-context learning demonstration retriever. We report the statistics of dataset in Table 5.

A.2 Training settings in MeCO

We finetune CLIP-ViT-Large with learning rate of 1e-4 and batch size of 256, ColQwen with LoRA tuning and a rank of 32, VLM2Vec with LoRA tuning and a rank of 16. The learning rate and batch size are respectively 1e-4 and 64 for both VLM2Vec and ColQwen. The experiments are conducted on Nvidia A800 80G.

B Case Study

B.1 Multimodal In-Context Learning Demonstration Retrieval for Different Tasks

We give four examples of multimodal in-context learning demonstrations for respectively VQA, Hateful Memes, Visual Caption, and Visual Caption Restorations. We note that for modality-independent retrievers, when the text input is only the task description it is the same for the query and doc and thus can not be used to demonstration retrieval.

B.2 Attention Map of two Retrievers

We give more examples showcasing that CLIP-ViT retriever trained on VCR-en and VCR-zh learns spurious correlations by incorrectly attributing importance to white areas, instead of (which happen to coincide with character regions). In contrast, VLM2Vec, a modality-integrated retriever can successfully attribute more attention to the true character areas and retrieval better in-context learning examples for multimodal large language models.

split	**VizWiz**	**VQAv2**	**OK-VQA**	**VCR-en**	**VCR-zh**	**Flickr30k**	**Hateful**
support	20523	—	—	100k	—	130k	6744
test	4319	2500	70.8	500	500	5000	2408

Table 5: Statistics of the datasets.

Figure 7: Cases of multimodal demonstration retrieval in different tasks.

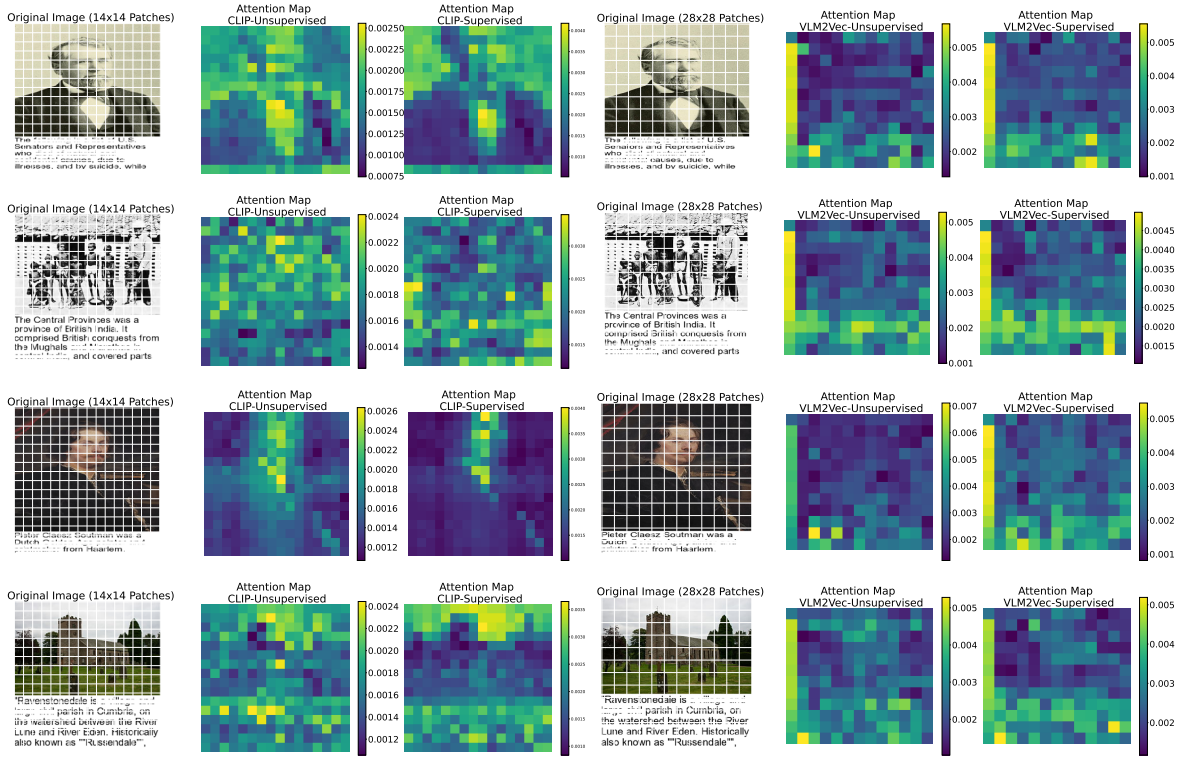

Figure 8: The attention map of image tokens to the pooling tokens. We use the attention scores in the first head of the first layer.