Re-M: Adapting Multi-Modal Large Language Models for Zero-Shot Cross-Modal Hybrid Retrieval and Reranking

Anonymous ACL submission

Abstract

Multi-Modal Large Language Models (MLLMs) demonstrate remarkable capabilities in vision-language understanding. Leveraging MLLMs to extract features for modern retrieval pipelines, including sparse retrieval, dense retrieval, and reranking, offers a promising direction that eliminates the need for expensive training data. In this paper, we investigate the feasibility of extracting high-quality sparse representations from MLLMs and propose a multi-perspective prompting method to enhance representational expressivity. Furthermore, we identify a significant performance disparity between image-to-text and text-to-image tasks during the reranking phase, indicating the necessity for distinct strategies. Building on these insights, we introduce Re-M, a two-stage zero-shot cross-modal retrieval framework. By integrating sparse-dense hybrid retrieval with asymmetric reranking, Re-M achieves performance that rivals or even surpasses retrieval-oriented dense and sparse baselines in zero-shot settings.

1 Introduction

Cross-modal retrieval (Wang et al., 2016) plays a pivotal role in diverse real-world applications, such as recommendation systems and search engines. With the advent of Multi-modal Large Language Models (MLLMs), researchers have begun adapting these models for retrieval tasks to leverage their superior comprehension capabilities. However, existing works (Faysse et al., 2025; Ma et al., 2024; Zhang et al., 2024) primarily utilize MLLMs as backbones for supervised contrastive fine-tuning, a process that necessitates a vast amount of high-quality labeled data. In contrast, MLLMs inherently possess extensive cross-modal knowledge derived from pre-training, achieving remarkable performance in understanding tasks like zero-shot image classification (Yan et al., 2024; Hong et al., 2025; Atabuzzaman et al., 2025). Consequently, harnessing the capabilities of MLLMs across various stages of modern retrieval pipelines represents a highly promising avenue for research.

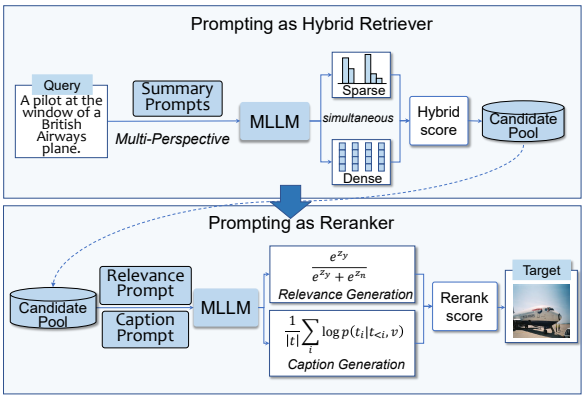


Figure 1: The proposed two-stage zero-shot retrieval framework, Re-M. The framework achieves hybrid retrieval with multi-perspectives prompting in the first stage and reranks the candidates via two different generation paradigms in the second stage.

Existing research on zero-shot retrieval has primarily focused on text modality, which demonstrate that prompting strategies enable LLMs to function as both retrievers and rerankers. For retrieval, standard approaches extract dense representations from the last hidden states using contrastive learning (Xu et al., 2024; Li et al., 2024b; Jiang et al., 2024c). While recent methods, such as PromptEOL (Jiang et al., 2024a), enhance discriminative power via Explicit One-word Limitation (EOL), they require resource-intensive training. To address this, PromptReps (Zhuang et al., 2024b) introduces a training-free hybrid retriever that generates dense and sparse representations in a single pass. Regarding reranking, LLM-based point-wise rerankers (Liu et al., 2024b; Long et al., 2025), pair-wise rerankers (Li et al., 2025; Qin et al., 2024), and list-wise rerankers (Ma et al., 2023; Chen et al., 2025) have been extensively ex-

plored. In comparison, exploration of the zero-shot setting in cross-modal retrieval and reranking has been quite limited. Only a few successful experiences (Jiang et al., 2024b) from text modality have been adapted to cross-modal domains and demonstrated their effectiveness.

To address the aforementioned research gap, this paper explores adapting MLLMs as hybrid retrievers and rerankers. Specifically, our research proceeds in two stages:

First, regarding hybrid retrieval, we evaluate two effective, training-free sparse representation methods from text-modality studies. Empirically, neither achieves satisfactory sparse and hybrid retrieval performance in cross-modal settings, primarily because they fail to incorporate prior dataset knowledge from diverse perspectives. To address this, we propose Multi-Perspective Prompting (MPP) sparsification strategy. MPP guides the MLLM to summarize information from dataset-specific angles and aggregates weights from these perspectives, yielding sparse representations with enhanced expressive power.

Second, for reranking, we evaluate two pointwise paradigms suitable for MLLM's context limits: Relevance Generation Paradigm (RGP) and Caption Generation Paradigm (CGP). RGP predicts text-image relevance directly, while CGP predicts candidate conditional probability. Interestingly, CGP excels in text-to-image reranking, whereas RGP performs better in image-to-text tasks. We attribute CGP's shortcomings in image-to-text reranking to deviations in the prior distribution of text candidates. Consequently, to avoid the high computational cost of debiasing, we implement distinct strategies for each direction, simultaneously optimizing both image-to-text and text-to-image retrieval.

Based on our analysis, we propose Re-M, a two-stage zero-shot cross-modal retrieval framework (Figure 1). First, Re-M enhances dense representations with MPP-derived sparse features to achieve robust hybrid retrieval. Second, it employs RCP and CGP reranking to optimize both text-to-image and image-to-text tasks. Extensive experiments show that Re-M rivals or surpasses specialized dense and sparse baselines in zero-shot settings. Ultimately, Re-M helps the community better understand the retrieval potential of MLLMs, enabling fine-tuning improvements with a more model-adaptive approach or utilizing it as a higher-quality distillation supervision signal.

2 Background

2.1 MLLM for Zero-Shot Dense Retrieval

Previous works, such as PromptEOL (Jiang et al., 2024a) and E5-V (Jiang et al., 2024b), show that Explicit One-word Limitation (EOL) embedding prompts can produce discriminative representations. Specifically, the MLLM F can be divided into the language modeling head g and other parts f containing the vision encoder, the projector, and the LLM transformer. which can be expressed as $F = g \circ f$. To obtain dense representations, we prepare prompt templates *<sent>\n Summary above sentence in one word:* for texts and *<image>\n Summary above image in one word:* for images. For image v, the MLLM image encoder is used to project v into the input embedding space, and the projected patch embeddings are filled into the template to replace *<image>*. For text t, it is directly replaced *<sent>* in the template. We uniformly define the filled sequence as $s = \{s_1, s_2, ..., s_l\}$, where l is the sequence length, then, the d-dimensional hidden states from the last layer can be expressed as

$$\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_l] = f(s) \quad (1)$$

where $\mathbf{H} \in \mathbb{R}^{d \times l}$ and $\mathbf{h}_i \in \mathbb{R}^{d \times 1}$. Guided by the templates, the semantics of images or text is aggregated at the final position $\mathbf{h}_l$, and thus can be regarded as representations for dense retrieval. For convenience, we define the text representation as $\mathbf{h}^{(t)}$ and the image representation as $\mathbf{h}^{(v)}$.

2.2 LLM for Zero-Shot Hybrid Retrieval

Sparse Representation With the language modeling head g, the hideen state can be mapped into logit $\mathbf{w} = g(\mathbf{h}_l) \in \mathbb{R}^{|\mathbb{V}|}$, where $|\mathbb{V}|$ denotes the vocabulary size. PromptReps (Zhuang et al., 2024b) shows $\mathbf{w}$ can be used for sparse representation through post-processing. Specifically, following the recipe of SPLADE (Formal et al., 2021), the initial sparse representation can be calculated as

$$\hat{\mathbf{w}} = \text{Round}(100 \times \log(1 + \text{ReLU}(\mathbf{w}))) \quad (2)$$

where the ReLU function avoids negative values, log-saturation prevents the domination of certain tokens, and the Round function quantizes logit values to the nearest integers. To build sparser representations, two potential methods from previous works are available:

- **Source-guided Sparse Representing (SSR)** *retains the tokens that appear in the original text while zero the other positions.* (Zhuang et al., 2024b)

$$\hat{\mathbf{w}}_{\text{ssr}} = \hat{\mathbf{w}} \odot (\mathbb{1}(\mathbb{V}_j \in t))_{j=1...|\mathbb{V}|}, \quad (3)$$

where t is the original text and $\mathbb{V}_j$ is the token corresponding to j^{th} indice in $\hat{\mathbf{w}}$.

- **Prediction-based Top-k Truncation (PTT)** *retains the tokens with the top-k largest logit values.* (Nie et al., 2025)

$$\hat{\mathbf{w}}_{\text{ptt}} = \hat{\mathbf{w}} \odot (\mathbb{1}(\hat{\mathbf{w}}_j \in \text{Topk}(\hat{\mathbf{w}})))_{j=1...|\mathbb{V}|}, \quad (4)$$

where $\hat{\mathbf{w}}_j$ is the token weight in j^{th} indice.

Hybrid Scoring For hybrid retrieval, a linear interpolation with appropriate weight α is introduced to fusion the dense and sparse scores score_d and score_s to generate hybrid scores score_h. Given the query-document dense representation $\mathbf{h}^{(q)}$ and $\mathbf{h}^{(d)}$ and sparse representation $\hat{\mathbf{w}}^{(q)}$ and $\hat{\mathbf{w}}^{(d)}$, the hybrid score is calculated as

$$\begin{aligned} \text{score}_\text{d} &= \mathbf{h}^{(q)} \cdot \mathbf{h}^{(d)} \quad \text{score}_\text{s} = \hat{\mathbf{w}}^{(q)} \cdot \hat{\mathbf{w}}^{(d)} \\ \text{score}_\text{h} &= \alpha\Gamma(\text{score}_\text{d}) + (1-\alpha)\Gamma(\text{score}_\text{s}) \end{aligned} \quad (5)$$

where α is the weight hyper-parameter, Γ is the min-max normalization. Since hidden states and logits can be obtained in a single forward pass, we integrate three distinct basic paradigms for zero-shot cross-modal retrieval within a hybrid retriever without double-counting.

2.3 LLM for Zero-Shot Reranking

LLMs for zero-shot reranking have been extensively studied, including point-wise (Liu et al., 2024b; Long et al., 2025), pair-wise (Li et al., 2025; Qin et al., 2024), and list-wise (Ma et al., 2023; Chen et al., 2025). Considering the context length limitations of MLLMs, we primarily focus on point-wise strategies for reranking. Point-wise requires a query and a candidate as input, prompts the LLM to measure the relevance, and outputs a ranking score. Point-wise methods include two paradigms: query generation and relevance generation. Query generation calculates the average log-likelihood of the query conditioned on the candidate as the ranking score (Sachan et al., 2022; Zhuang et al., 2023). Relevance generation applies the softmax function on the logit of token 'yes' and 'no' or other relevance levels (Zhuang et al., 2024a; Liu et al., 2024b) to generate ranking scores.

3 Methodology

3.1 Acquisition of Sparse Representation

Earlier, we detailed constructing hybrid retrievers via two sparse methods, previously limited to document retrieval. We first explore their cross-modal feasibility. Since images lack original sentences, only SSR is applicable for the image modality. For the text modality, we build corresponding hybrid retrievers to compare both strategies, using E5-V prompts and setting PTT with k=128. Results on Flickr30K are shown in Figure 2. We report average recall@1 on Flickr30K (Lin et al., 2014) in Figure 2. Notably, similar conclusions hold for other datasets (e.g., MSCOCO (Plummer et al., 2015)), with details in Appendix C.

As shown in Figure 2, there is a contradiction between sparse and hybrid retrieval with SSR and PTT, which indicates that neither of them can perform well simultaneously. We conjecture that, compared to SSR, PTT's sparse representations are simply the spatial projection of dense representations; therefore, they share more identical tokens (blue tokens in Table 1) when their corresponding dense representations are semantically similar. This property makes PTT's sparse representations suitable for token matching, leading to respectable sparse retrieval performance. However, this results in dense and sparse representations encoding homogeneous information. For hybrid retrieval to outperform dense-only or sparse-only approaches, the two representations should provide heterogeneous content. Relying solely on PTT for both modalities thus limits the potential gains in hybrid retrieval. Therefore, while each sparsification method has its strengths, further research is needed to develop new sparse approaches better suited for cross-modal retrieval and improve both sparse and hybrid retrieval performance simultaneously.

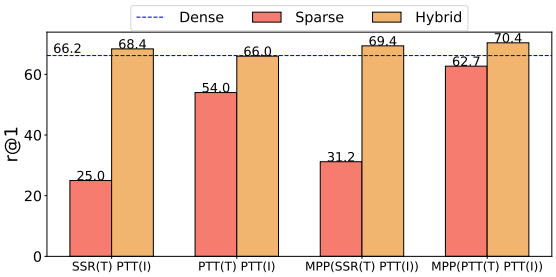


Figure 2: Average recall@1 on Flickr30K of text-to-image and image-to-text on LLaVA-Next-7B. The letter in parentheses denotes the modality, where "I" means image modality and "T" means text modality.

Image & Text	Method	Modality	Perspective	Top 10 Predicted Tokens
 A brown dog is shown standing in the water near a muddy beach.	SSR	text	coarse	_Dog, _Beach, _Mud, _Brown, _Standing, _Water, _Near, _Shown, _Dy
	PTT	text	coarse	_Dog , _Beach , _Brown , _Be, _W , _A, _Sw, _Water, _Yes, _The
		image	coarse	_F, _Dog , _Mist, _M, _Rain, _Beach , _W , _Cloud, _H,, _D
	MPP	text	person or objects	_Dog , _Beach , _M, _Brown , _, _Water , _Sand, _Be, _One, _No
			relations	_Dog , _Beach , _M, _Water , _Sand, _None, _Own, _Near, _No, _Brown
			environment, weather or places	_Beach , _Water , _Sh, _M, _Sw, _Ocean, _W, _River, _Be, _Coast
			actions or movements	_Stand, _Dog , _Sw, _St, _W, _Beach , _M, _Wait, _Walk, _Dr
			appearance	_Dog , _Brown , _M, _Color, _Water , _The, _, _Beach , _A, _Be
		image	person or objects	_Dog , _F, _, _None, _No, _One, _Water , _Nothing, _Cloud, _Sand
			relations	_Dog , _None, _No, _, _F, _Un, _Run, _Water , _Sep, _One
			environment, weather or places	_F, _Mist, _M, _Beach , _Cloud, _Sw, _W, _Rain, _D, _H
			actions or movements	_Walk, _Run, _Running, _Dog , _Stand , _J, _Sw, _W, _Play, _S
			appearance	_Dog , _Brown , _F, _M, _W, _Water , _D, _Gray, _Mist, _Sand

Table 1: The case study of the top 10 predicted tokens with SSR, PTT, and MPP. The 'coarse' indicates the ordinary EOL templates used in the Background Section.

3.2 Multi-Perspectives Prompting

Motivation When analyzing the sparse weights from existing methods, we find that MLLM struggles to grasp the aspects that summarize information in the dataset and focuses on core tokens via coarse-grained summary prompt templates. For example, when summarizing the image, predicted tokens with the top-10 largest weights exclude token '_Standing', which reflects the action of the dog, and token '_Brown' (red tokens in Table 1) reflecting the color of the dog. Meanwhile, the empirical success of test-time scaling in other areas, such as semantic text similarity (Lei et al., 2024) and image classification (Xiao, 2024), motivates us to consider predefined dataset-specific perspectives in prompt templates, enabling the MLLM to summarize information from multiple angles and thereby supplement the model with multifaceted prior knowledge about the dataset. Thus, we propose **M**ulti-**P**erspectives **P**rompting(MPP) to improve sparse representations.

Prompting Design Through analysis of the image-text pairs in datasets, we realize that although data annotators describe in diverse forms, they always annotate in a constrained number of dataset-specific perspectives. For example, images and texts in MSCOCO and Flickr30K can be described from five angles: person or object, relations, environments, actions, and appearances. Based on this discovery, we collect a set of perspectives for each dataset [1]. We revise prompt templates to *<image>/<sent>\n Summary the <angle> in the above image/sentence in one word:*. Then, we replace the *<angle>* token with a perspective from the set and obtain a group of prompt templates $\mathcal{T}$. We denote $L = |\mathcal{T}|$ as the sum of perspectives to guide MLLM in summarizing information from various perspectives. Concrete modified prompt templates for different benchmarks are shown in the Appendix B. We quantize token weights for each prompt template and accumulate them to produce sparse representations $\hat{\mathbf{w}}_{\text{mpp}}$:

$$\hat{\mathbf{w}}_{\text{mpp}} = \sum_{a \in \mathcal{T}} \hat{\mathbf{w}}_a \tag{6}$$

where a is a prompt template with one perspective in $\mathcal{T}$, $\hat{\mathbf{w}}_a$ is the sparse representation produced by the prompt template a. We apply PTT to both modalities for sparsity, enabling the MLLM to explicitly capture the salient perspectives in the dataset content while maintaining high token matching accuracy.

Efficiency Optimization During retrieval, we reserve the prompt templates used by basic retrievers for dense representations to embed global semantics and use MPP for sparse representations. To address inefficient multiple forward passes, we introduce prompt-wise attention masking (Kim and Angelova, 2025; Li et al., 2023) and concatenate all prompt templates into a single sequence, as shown in Figure 3. We adjust the attention mask to prevent subsequent prompts from attending to earlier ones and extract representations at the end of each prompt to simulate the last pooling operation. In this way, MPP reduces latency while retaining the advantage of acquiring multiple retrieval paradigms within one forward pass and achieves simultaneous

[1] In this paper, we adopt a manual approach for perspective construction, reserving the method of iterative optimization via LLMs for future work.

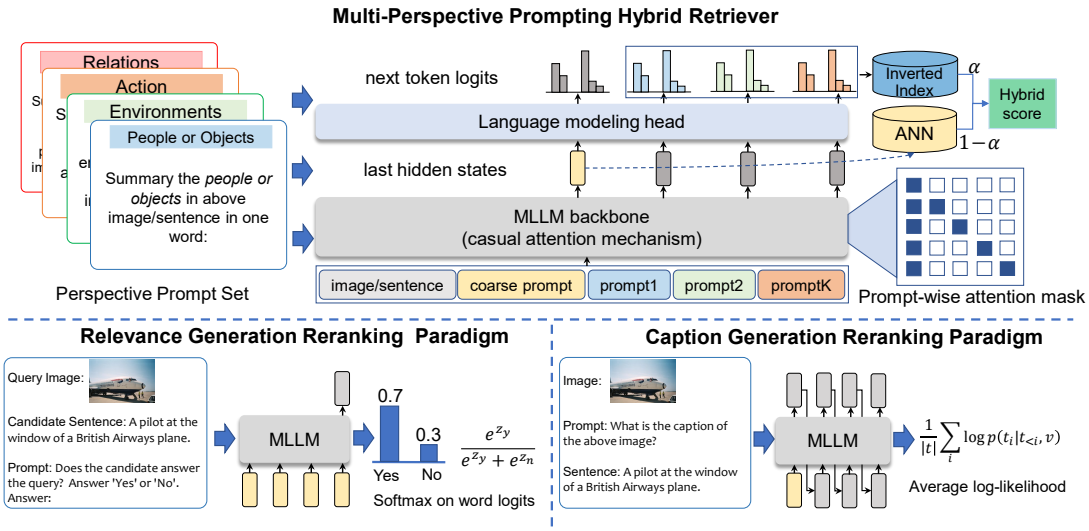


Figure 3: The illustration of our two-stage zero-shot cross-model framework, Re-M. In the retrieval phase, we propose multi-perspective prompting to enhance the performance of sparse retrieval and hybrid retrieval. In the reranking phase, we employed relevance generation for text reranking and caption generation for image reranking to improve performance on both sides simultaneously.

progress in sparse and hybrid retrieval.

Analysis of MPP From Table 1, we find MPP effectively addresses the issue of SSR and PTT with a coarse summary prompt. MLLM understands how to summarize information from various angles and predict the key tokens (highlighted green) with a top-weight ranking, such as the tokens '_Brown' and '_Stand'. Finally, it is intuitive that some key tokens appear in various predictions repeatedly, and the weight accumulation mechanism emphasizes the key tokens (highlighted orange) with a larger weight to enable the sparse representations to be more discriminative, avoiding interference from similar negative candidates. Additionally, we report the performance of MPP based on SSR and PTT in Figure 2, denoted as "MPP(SSR(T), PTT(I))" and "MPP(PTT(T), PTT(I))". It is observed that when achieving similar hybrid retrieval results, MPP based on PTT gains remarkable sparse retrieval performance compared to MPP based on SSR. Therefore, **we employ construct MPP upon PTT by default to support hybrid retrieval.**

3.3 Prompting MLLM as a Reranker

In the modern retrieval pipeline, the final performance can be significantly improved by employing a reranker that follows the retriever. However, currently, there is no research to prompt MLLM as a reranker in cross-modal retrieval. Thus, we adapt two existing paradigms to enable MLLMs to acquire reranking capability.

Relevance Generation Paradigm (RGP) Given an image v and a text t, the MLLM is prompted with a template to determine whether they are relevant and outputs "yes" or "no". Denote the logit of token 'Yes' and token 'No' with z_Y and z_N, then the ranking score $r(t, v)$ is calculated by applying a softmax function:

$$r(t, v) = \frac{\exp(z_Y)}{\exp(z_Y) + \exp(z_N)}. \quad (7)$$

There are numerous prompt templates for selection. We referenced the templates provided in a systematic study on document reranking (Sun et al., 2025) and adapted them for image-text reranking. Preliminary experiments were conducted to select the optimal prompt template from available options. We report the results for all templates in Appendix D.

Caption Generation Paradigm (CGP) Given an image v and a text t, the log-likelihood of the sentence conditioned on the matching image is calculated as the reranking scores. Normally, we use $\log p(t|v)$ to estimate the image-to-text reranking score and $\log p(v|t)$ to calculate the text-to-image reranking score. Assume that the prior distribution $p(v)$ is uniform, according to Bayes' rule, $\log p(v|t)$ should be proportional to $\log p(t|v)$. Therefore, we prompt MLLM to calculate the average log-likelihood $\log p(t|v)$ as the reranking score

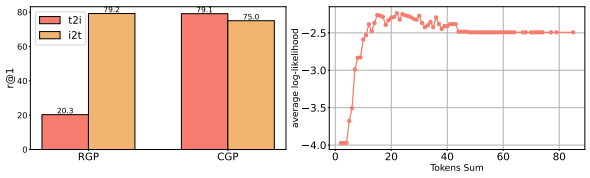


Figure 4: The recall@1 results of RGP and CGP (left) and average log-likelihood generating texts conditioned on no image (right) on Flickr30K.

$r(t, v)$ for both retrieval settings:

$$r(t, v) = \frac{1}{|t|} \sum_{i}^{|t|} \log p(t_i | t_{<i}, v, \theta) \quad (8)$$

where θ denotes the parameters of MLLM and $|t|$ is the length of text tokens.

Analysis of Performance Disparity We report our results on the Flickr30K validation set in the left of Figure 4, which demonstrates that the optimal methods differ between text-to-image and image-to-text tasks. Specifically, CGP demonstrates significantly better performance in text-to-image reranking, while RGP holds an advantage in image-to-text reranking. Considering MLLM as a language model fundamentally, we believe that the CGP paradigm aligns more closely with MLLM training patterns than RGP, which should yield superior results on both sides. However, as shown in the right of Figure 4, we observed significant discrepancies in the inherent generation probabilities across texts of varying lengths. When presented with one image and multiple candidate texts $(i, t_1, \ldots, t_N)$, the probability $p(t_j|i)$ becomes distorted by the influence of $p(t_j)$'s own probability, introducing bias. In contrast, when presented with one image and multiple candidate texts $(t, i_1, \ldots, i_N)$, the fixed variable t in $p(i_j|t) \propto p(t|i_j)$ eliminates this interference. To address this issue in graph-to-text reordering, two forward passes are required: calculating $p(t_i|i)$ and $p(t_i)$, followed by bias removal. However, this approach significantly increases computational complexity. Therefore, during reranking, **we select RGP for image-to-text and CGP for text-to-image to ensure optimal performance on both sides while maintaining efficiency.**

3.4 Retrieval Pipeline: Re-M

As shown in Figure 3, we design the Re-M pipeline for training-free, zero-shot cross-modal retrieval. In the first stage, we exploit dense representations for semantic similarity calculation with ANN and sparse representations for token matching with an inverted index separately. The hybrid score calculation method for text and images is consistent with Equation 5, except that the sparse representation weights are replaced by the multi-perspectives prompting strategy (Equation 6). In the second stage, we select the top N candidates with the highest hybrid scores and rerank the candidate texts with RGP, while the candidate images with CGP.

4 Experiment

4.1 Experiment Setup

Dataset and Evaluation We assess baselines and Re-M on widely used image-text retrieval datasets MSCOCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015) following the Karpathy split (Karpathy and Fei-Fei, 2015). Each image is combined with five sentences. To evaluate the generalizability of our model on domain datasets, we select the CUHK-PEDES (Li et al., 2017), ICFG-PEDES (Ding et al., 2021), and RSTPReid (Zhu et al., 2021) benchmarks in text-based person retrieval. We report recall@k with k=1, 5, 10 for all retrieval datasets.

Baselines We compare Re-M with widely recognized dense models for zero-shot image-text retrieval. We select CLIP with ViT-B and ViT-L (Radford et al., 2021), BLIP with ViT-L (Li et al., 2022), EVA-02-CLIP with 5B parameters (Sun et al., 2023) and E5-V (Jiang et al., 2024b). We also introduce some image lexical models, such as LexLIP (Luo et al., 2023), STAIR (Chen et al., 2023), and D2S (Nguyen et al., 2024), as baselines.

Implementation Details We design the MPP retriever and Re-M pipeline based on two strong baseline MLLM: LLaVA-Next-8B (Li et al., 2024a) and LLaVA-Next-7B (Liu et al., 2024a). We choose the Faiss (Douze et al., 2025) library to construct the ANN index with cosine similarity to rank and use brute force search. We use Pyserini (Lin et al., 2021) to build the inverted index for sparse retrieval. The specific hybrid method follows ranx.fuse (Bassani and Romelli, 2022). We set k to 30 for image-text retrieval of MPP, and to 10 for text-based person retrieval. Due to the existence of repeated tokens, this can ensure the length of MPP is shorter than PTT. In the reranking stage, we set the candidate size N to 5 and evaluate the reranker on the recall@1 metric. During the evaluation, we use the

Method		MSCOCO						Flickr30K					
		t2i			i2t			t2i			i2t		
		r@1	r@5	r@10	r@1	r@5	r@10	r@1	r@5	r@10	r@1	r@5	r@10
Dense Baselines	CLIP ViT-B	30.5	56.0	66.8	51.0	74.9	83.5	58.8	83.3	89.8	77.8	95.0	98.2
	CLIP ViT-L	37.0	61.6	71.5	58.1	81.1	87.8	67.3	89.0	93.3	87.2	98.3	99.4
	BLIP ViT-L	48.4	74.4	83.2	63.5	86.5	92.5	70.0	91.2	95.2	75.5	95.1	97.7
	EVA-02-CLIP	51.1	75.0	82.7	**68.8**	87.8	92.8	78.8	94.2	96.8	**93.9**	99.4	99.8
	E5-V	52.0	76.5	84.7	62.0	83.6	89.7	79.5	95.0	97.6	88.2	98.7	99.4
Sparse Baselines	LexLIP	53.2	79.1	86.7	70.2	90.7	95.2	78.4	94.6	97.1	91.4	99.2	99.7
	STAIR	41.1	65.4	75.0	57.7	80.5	87.3	66.6	88.7	93.5	81.2	96.1	98.4
	D2S	54.5	80.6	-	-	-	-	79.8	95.9	-	-	-	-
LLaVA-Next-LLaMA-8B	Dense	34.6	60.5	71.2	43.3	68.6	78.2	60.1	83.7	90.3	72.4	90.6	95.6
	MPP Sparse	25.8	51.4	63.3	29.6	57.7	70.4	52.5	79.4	86.4	55.6	83.6	90.4
	MPP Hybrid	37.3	63.7	74.1	46.0	71.2	80.5	64.4	86.7	91.9	74.1	91.9	96.1
	Re-M	51.6	-	-	50.7	-	-	79.1	-	-	79.2	-	-
LLaVA-Next-Mistral-7B	Dense	34.3	59.7	70.3	42.7	67.2	77.1	60.5	83.0	89.0	72.3	90.9	94.6
	MPP Sparse	30.5	56.1	67.5	39.2	66.8	77.7	57.0	82.3	89.1	68.4	89.6	94.5
	MPP Hybrid	38.1	64.5	74.8	46.1	72.4	81.5	65.5	87.5	92.9	75.4	93.9	97.3
	Re-M	**52.1**	-	-	49.4	-	-	**79.8**	-	-	80.0	-	-

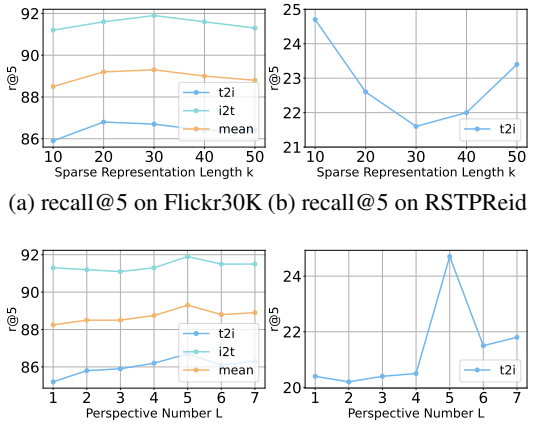
Table 2: The image-text retrieval results on MSCOCO and Flickr30K. The models painted grey are finetuned on MSCOCO and Flickr30K, and the others are evaluated in the zero-shot setting. Here, the letter 'r' indicates recall.

validation set to determine the optimal weights α (ranging from 0.1 to 0.9) for the hybrid score. All experiments are run on two Nvidia V100 GPUs.

4.2 Main Results

Image-Text Retrieval We present the image-text retrieval results of Re-M and baselines on MSCOCO and Flickr30K in Table 2. The first observation is that MPP hybrid retrieval outperforms both dense and sparse retrieval alone, confirming the effectiveness of the weighted hybrid method in cross-modal retrieval. Recalls are improved by 3 to 4 points, and for r@1 with LLaVA-Next-7B, the improvement reaches 5 points. Another finding is that the Re-M reranking performance outperforms strong zero-shot dense baselines. Meanwhile, the reranking pipeline produces competitive results compared to finetuned sparse baselines. Although exceeding EVA-02-CLIP is difficult, our Re-M pipeline maintains its entire generation capability, representing a fundamental advantage.

Text-based Person Retrieval To evaluate the generalization of Re-M, we report results on text-based person retrieval in Table 3, which can be viewed as a domain task. In Table 3, the results demonstrate that (1) MPP hybrid retrievers are still effective on domain tasks. (2) Re-M pipeline can outperform strong dense baselines on TBPR tasks, maintaining similar phenomena with ITR. The r@1 results after reranking are even closer to r@5 of hybrid retrieval. This suggests the generalization of our method without finetuning on domain datasets.



(a) recall@5 on Flickr30K (b) recall@5 on RSTPReid

(c) recall@5 on Flickr30K (d) recall@5 on RSTPReid

Figure 5: Sparse representation length k (Figure (a) and (b)) and perspective number L (Figure (c) and (d)) of MPP analysis on Flickr30K and RSTPReid, reporting the r@5 of MPP hybrid retrieval on LLaVA-Next-8B.

4.3 Ablation Study

Analysis of Sparse Representation Length k Figure 5a and 5b present the trend of r@5 for MPP hybrid retrieval on LLaVA-Next-8B across the Flickr30K and RSTPReid datasets, with the increase of length k corresponding to that of the sparse representations w_a from each perspective. In Figure 5a, r@5 attains a high value even at a

Method		CUHK-PEDES			ICFG-PEDES			RSTPReid		
		r@1	r@5	r@10	r@1	r@5	r@10	r@1	r@5	r@10
Dense Baselines	CLIP ViT-B	5.9	14.7	21.0	3.2	9.3	13.9	8.9	21.1	31.2
	CLIP ViT-L	10.1	22.8	31.2	5.2	13.3	19.4	11.2	26.4	37.8
	BLIP ViT-L	11.2	26.4	37.8	8.1	19.0	26.1	16.9	38.0	50.9
	EVA-02-CLIP	**29.3**	47.6	56.4	**18.0**	33.6	42.2	**27.0**	50.4	62.0
	E5-V	20.6	37.5	47.0	7.5	17.6	24.3	21.6	42.4	54.8
LLaVA-Next-LLaMA-8B	Dense	5.3	12.7	16.9	1.6	5.0	7.6	7.6	20.7	31.6
	MPP Sparse	5.8	13.3	19.0	1.6	5.5	8.7	6.3	21.3	32.2
	MPP Hybrid	7.0	15.4	21.5	2.2	6.8	10.6	8.6	24.7	36.1
	Re-M	14.9	-	-	5.7	-	-	20.0	-	-
LLaVA-Next-Mistral-7B	Dense	6.8	15.9	22.4	2.3	6.8	10.2	9.6	26.7	36.8
	MPP Sparse	11.7	25.0	32.8	4.8	12.3	17.5	13.4	31.9	43.0
	MPP Hybrid	12.9	26.7	35.7	5.0	12.9	18.4	14.8	34.1	16.9
	Re-M	24.8	-	-	10.7	-	-	26.8	-	-

Table 3: The text-based person retrieval results on CUHK-PEDES, ICFG-PEDES, and RSTPReid. In the table, the letter 'r' indicates recall.

length of 10. As k increases, recall improves gradually and peaks when the length reaches 30. On the text-based person retrieval task (RSTPReid, Figure 5b), the best performance is observed at length 10, with degradation at longer lengths. These results suggest that high sparsity can be maintained while ensuring effective hybrid retrieval.

Analysis of Perspective Number L Figure 5c and 5d report the trend of r@5 for MPP hybrid retrieval on LLaVA-Next-8B across the Flickr30K and RSTPReid datasets under varying numbers L of perspective prompts. The results show that hybrid performance peaks when the number of perspectives reaches five, at which point the perspective set sufficiently covers the dataset content. However, introducing less relevant aspects leads to performance degradation. This indicates that including only the most pertinent perspectives is critical to produce reliable sparse representations, while irrelevant ones should be excluded.

4.4 Discussion of the Three-stage Pipeline

Our evaluation shows that sparse-only MPP retrieval achieves high recall when the retrieval set is sufficiently large. This inspires the conception of a three-stage pipeline that first applies MPP sparse retrieval to select 200 candidates coarsely, followed by hybrid retrieval and reranking. We analyze the time-cost upper bound of this pipeline on the LLaVA-Next-8B, assuming all tokens in the sparse representations are matched, and the maximum length is 150 (30 tokens per perspective). The pipeline is compared to CLIP ViT-L, and FLOP trends are illustrated in Figure 6 as the data scale increases. When the data scale exceeds about 125 billion, the three-stage pipeline has lower FLOPs than CLIP, indicating its feasibility for super-large-scale search. It is worth noting that Figure 6 reflects the worst-case situation. Through the use of sparser representations, high-performance GPUs, and the KV-cache technique, the time cost can be further reduced for applications on large-scale data.

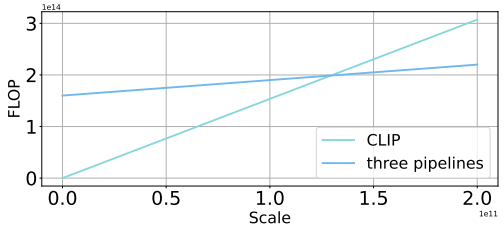


Figure 6: FLOP of three-stage pipeline and CLIP.

5 Conclusion

In this work, we investigate the Re-M pipeline for adapting MLLM to zero-shot cross-modal retrieval and reranking via prompting. We design an efficient hybrid retriever and propose Multi-Perspective Prompting to improve sparse and hybrid retrieval simultaneously. Meanwhile, we introduce two paradigms to prompt MLLM as the reranker to address the asymmetric improvements. Experiments on image-text retrieval datasets and domain text-based person retrieval benchmarks demonstrate the effectiveness of Re-M.

Limitations

- When designing perspectives for datasets, we must investigate the concrete images and sentences in the dataset and summarize the description angles of annotators. This process is not automatic, and humans struggle to produce the optimal dataset-specific perspectives manually. However, we did not find a proper method to generate summary perspectives without humankind. For example, we should have considered prompting the LLM to generate and iteratively optimize perspectives, like producing schemas for knowledge graphs automatically. We believe this should be researched in the future.

- Because of the time limitation, we only set the size of the candidate pool for re-ranking to 5 and report r@1 in our experiments. This setting constrains the performance of RGP and CGP, so we plan to supplement results to set the size of the candidate pool to 20 and record results on r@1, r@5, and r@10 metrics to further elaborate on the effectiveness of our re-ranking paradigms.

- In our studies, to achieve improvements on both retrieval settings, we adopt RGP on image-to-text retrieval and CGP on text-to-image retrieval. However, this operation indicates that we do not explain why RGP and CGP cannot perform well on both retrieval settings at the same time. In the paper, we simply report the average log-likelihood of various token sequence lengths of sentences conditioned on no images. This may modify the CGP log-likelihood of image-to-text retrieval and improve the results. But there is no experimental evidence to support this. Besides, this paper lacks an analysis of RGP.

- Due to the absence of an efficiency optimization mechanism, the process of re-ranking is computationally intensive and leads to huge FLOPs. This influences the application of Re-M on small-scale data. In the future, researchers should propose an approach to solve this issue.

References

Md Atabuzzaman, Andrew Zhang, and Chris Thomas. 2025. Zero-shot fine-grained image classification using large vision-language models. *arXiv preprint arXiv:2510.03903*.

Elias Bassani and Luca Romelli. 2022. ranx.fuse: A python library for metasearch. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 4808–4812. ACM.

Chen Chen, Bowen Zhang, Liangliang Cao, Jiguang Shen, Tom Gunter, Albin Jose, Alexander Toshev, Yantao Zheng, Jonathon Shlens, Ruoming Pang, and Yinfei Yang. 2023. STAIR: Learning sparse text and image representation in grounded tokens. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15079–15094, Singapore. Association for Computational Linguistics.

Shijie Chen, Bernal Jimenez Gutierrez, and Yu Su. 2025. Attention in large language models yields efficient zero-shot re-rankers. In *The Thirteenth International Conference on Learning Representations*.

Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. 2021. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data*.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *ICLR*.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

Yunqi Hong, Sohyun An, Andrew Bai, Neil YC Lin, and Cho-Jui Hsieh. 2025. Unlabeled data improves fine-grained image zero-shot classification with multimodal llms. *arXiv preprint arXiv:2506.03195*.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2024a. Scaling sentence embeddings with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3182–3196, Miami, Florida, USA. Association for Computational Linguistics.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024b. E5-v: Universal

embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*.

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

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dahun Kim and Anelia Angelova. 2025. Context-adaptive multi-prompt embedding with large language models for vision-language alignment. *arXiv preprint arXiv:2508.02762*.

Yibin Lei, Di Wu, Tianyi Zhou, Tao Shen, Yu Cao, Chongyang Tao, and Andrew Yates. 2024. Meta-task prompting elicits embeddings from large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10141–10157, Bangkok, Thailand. Association for Computational Linguistics.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023. Making large language models a better foundation for dense retrieval. *arXiv preprint arXiv:2312.15503*.

Chaofan Li, Zheng Liu, Shitao Xiao, Yingxia Shao, and Defu Lian. 2024b. Llama2vec: Unsupervised adaptation of large language models for dense retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3490–3500.

Jieran Li, Xiuyuan Hu, Yang Zhao, Shengyao Zhuang, and Hao Zhang. 2025. Leveraging reference documents for zero-shot ranking via large language models. *arXiv preprint arXiv:2506.11452*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2356–2362.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Wenhan Liu, Yutao Zhu, and Zhicheng Dou. 2024b. Demorank: Selecting effective demonstrations for large language models in ranking task. *arXiv preprint arXiv:2406.16332*.

Kehan Long, Shasha Li, Chen Xu, Jintao Tang, and Ting Wang. 2025. Precise zero-shot pointwise ranking with llms through post-aggregated global context information. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2384–2394.

Ziyang Luo, Pu Zhao, Can Xu, Xiubo Geng, Tao Shen, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Lexlip: Lexicon-bottlenecked language-image pre-training for large-scale image-text sparse retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11206–11217.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*.

Thong Nguyen, Mariya Hendriksen, Andrew Yates, and Maarten de Rijke. 2024. Multimodal learned sparse retrieval with probabilistic expansion control. In *European Conference on Information Retrieval*, pages 448–464. Springer.

Zhijie Nie, Richong Zhang, and Zhanyu Wu. 2025. A text is worth several tokens: Text embedding from LLMs secretly aligns well with the key tokens. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7683–7694, Vienna, Austria. Association for Computational Linguistics.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting

region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, and 1 others. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518.

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*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3781–3797, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Shuoqi Sun, Shengyao Zhuang, Shuai Wang, and Guido Zuccon. 2025. An investigation of prompt variations for zero-shot llm-based rankers. In *European Conference on Information Retrieval*, pages 185–201. Springer.

Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. 2016. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*.

Han Xiao. 2024. Scaling test-time compute for embedding models. Jina AI Blog.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May Dongmei Wang, Joyce C. Ho, Chao Zhang, and Carl Yang. 2024. BMRetriever: Tuning large language models as better biomedical text retrievers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22234–22254, Miami, Florida, USA. Association for Computational Linguistics.

Xudong Yan, Songhe Feng, Yang Zhang, Jian Yang, Yueguan Lin, and Haojun Fei. 2024. Leveraging mllm embeddings and attribute smoothing for compositional zero-shot learning. *arXiv preprint arXiv:2411.12584*.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*.

Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. 2021. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM international conference on multimedia*, pages 209–217.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2024a. Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 358–370, Mexico City, Mexico. Association for Computational Linguistics.

Shengyao Zhuang, Bing Liu, Bevan Koopman, and Guido Zuccon. 2023. Open-source large language models are strong zero-shot query likelihood models for document ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8807–8817, Singapore. Association for Computational Linguistics.

Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2024b. PromptReps: Prompting large language models to generate dense and sparse representations for zero-shot document retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4375–4391, Miami, Florida, USA. Association for Computational Linguistics.

A Dataset Statistics

We report the data split statistics in Table 4. For ICFG-PEDES, we use the train dataset as the validation dataset to select the best hybrid weight.

B Prompt Templates

For MSCOCO and Flickr30K datasets in image-text retrieval, we use the following perspective prompts to summarize information:

- <image>/<sent>\n Summary the people or objects in above image/sentence in one word:
- <image>/<sent>\n Summary the relations, such as belongings or spatial position, between main people or objects in above image/sentence in one word:
- <image>/<sent>\n Summary the environment, weather or places in above image/sentence in one word:

- <image>/<sent>\n Summary the actions or movements of main people or objects in above image/sentence in one word:

- <image>/<sent>\n Summary the appearance, such as color, material, decoration and so on, of main people or objects in above image/sentence in one word:

For CUHK-PEDEs, ICFG-PEDES, and RSTPReid datasets in text-based person retrieval, we use the following prompts to summarize information:

- <image>/<sent>\n Summary the gender of person in above image/sentence in one word:

- <image>/<sent>\n Summary the actions or movements of person in above image/sentence in one word:

- <image>/<sent>\n Summary the objects in above image/sentence in one word:

- <image>/<sent>\n Summary the wearing of person in above image/sentence in one word:

- <image>/<sent>\n Summary the appearance and decoration details of person, such as color, pattern and so on, in above image/sentence in one word:

For RGP, we exploit the following prompts to demand MLLM to generate relevance score.

When using RGP:

- text-to-image: For the following query sentence and candidate image, judge whether they are relevant. Output 'Yes' or 'No'.\n Query Sentence: <sent> Candidate Image: <image> Output:

- image-to-text: For the following query image and candidate sentence, judge whether they are relevant. Output 'Yes' or 'No'.\n Query Image: <image> Candidate Sentence: <sent> Output:

For CGP, we exploit the following prompts to demand MLLM to generate a relevance score:

- Image: <image>\n What is the caption of the above image? <sent>"

Dataset	train	val	test
MSCOCO	113200	5000	5000
Flickr30K	29800	1000	1000
CUHK-PEDES	34054	3078	3074
ICFG-PEDES	34674	0	19848
RSTPReid	18505	1000	1000

Table 4: Dataset statistics.

C Results of Basic SSR, PTT, and MPP

We report the whole retrieval results of SSR and PTT on dense, sparse, and hybrid retrieval in Table C. From this table, we can clearly see that simply utilizing SSR or PTT cannot achieve simultaneous improvements of sparse and hybrid retrieval. In addition, although MPP with SSR(T) and PTT(I) obtains similar hybrid results compared to MPP with PTT(T) and PTT(I), the sparse result is very low. Therefore, we adopt the MPP(PTT(T) PTT(I)) used in this paper. Furthermore, well-performing sparse results provide the possibility of a three-stage pipeline.

D Results of RGP Prompt Variations

Here we provide some other RGP prompt variations:

RGP_1:

- text-to-image: Given a candidate and a query, predict whether the candidate includes an answer to the query by producing either 'Yes' or 'No'.\n Candidate: <image>\n Query: <sent>\n Does the candidate answer the query? Answer:

- image-to-text: Given a candidate and a query, predict whether the candidate includes an answer to the query by producing either 'Yes' or 'No'.\n Candidate: <sent>\n Query: <image>\n Does the candidate answer the query? Answer:

RGP_2:

- text-to-image: Query: <sent>\n Candidate: <image>\n Does the candidate answer the query? Answer 'Yes' or 'No'. Answer:

- image-to-text: Query: <image>\n Candidate: <sent>\n Does the candidate answer the query? Answer 'Yes' or 'No'. Answer:

Method		MSCOCO						Flickr30K					
		t2i			i2t			t2i			i2t		
		r@1	r@5	r@10	r@1	r@5	r@10	r@1	r@5	r@10	r@1	r@5	r@10
LLaVA-Next-LLaMA-8B	Dense	34.6	60.5	71.2	43.3	68.6	78.2	60.1	83.7	90.3	72.4	90.6	95.6
	SSR(T) PTT(I) Sparse	17.2	37.5	48.3	5.5	18.8	29.8	30.4	56.7	68.4	12.3	32.4	44.6
	SSR(T) PTT(I) Hybrid	36.5	63.1	73.2	44.1	70.0	78.6	62.1	85.8	91.5	73.0	90.6	95.8
	PTT(T) PTT(I) Sparse	26.0	50.9	62.3	32.4	58.3	69.7	47.6	74.9	83.2	57.5	82.6	89.7
	PTT(T) PTT(I) Hybrid	34.4	60.3	71.0	43.0	68.2	78.2	59.8	83.3	90.3	72.0	90.4	95.2
	MPP(SSR(T) PTT(I)) Sparse	18.3	38.7	49.2	7.4	21.7	31.1	32.2	57.6	67.1	16.6	36.9	49.3
	MPP(SSR(T) PTT(I)) Hybrid	**38.6**	**65.1**	**75.3**	44.8	70.1	79.0	**64.6**	**87.4**	**92.3**	73.2	91.3	95.7
	MPP(PTT(T) PTT(I)) Sparse	25.8	51.4	63.3	29.6	57.7	70.4	52.5	79.4	86.4	55.6	83.6	90.4
	MPP(PTT(T) PTT(I)) Hybrid	37.3	63.7	74.1	**46.0**	**71.2**	**80.5**	64.4	86.7	91.9	**74.1**	**91.9**	**96.1**
LLaVA-Next-Mistral-7B	Dense	34.3	59.7	70.3	42.7	67.2	77.1	60.5	83.0	89.0	72.3	90.9	94.6
	SSR(T) PTT(I) Sparse	17.1	36.3	46.7	5.5	18.3	27.3	34.8	59.1	68.5	15.3	36.4	48.4
	SSR(T) PTT(I) Hybrid	35.9	61.6	72.0	42.9	67.7	77.7	63.1	85.4	90.6	73.7	91.0	95.1
	PTT(T) PTT(I) Sparse	25.4	49.2	60.7	31.3	56.0	66.9	48.6	74.8	81.9	59.4	83.4	90.7
	PTT(T) PTT(I) Hybrid	34.1	59.6	70.2	42.2	67.1	76.7	60.2	82.8	88.7	71.9	90.6	94.9
	MPP(SSR(T) PTT(I)) Sparse	18.5	37.6	47.5	12.4	30.6	41.8	34.2	58.9	68.0	28.2	51.9	63.3
	MPP(SSR(T) PTT(I)) Hybrid	**38.7**	**65.0**	**75.1**	43.8	68.8	79.2	65.2	87.2	92.1	73.6	90.6	95.6
	MPP(PTT(T) PTT(I)) Sparse	30.5	56.1	67.5	39.2	66.8	77.7	57.0	82.3	89.1	68.4	89.6	94.5
	MPP(PTT(T) PTT(I)) Hybrid	38.1	64.5	74.8	**46.1**	**72.4**	**81.5**	**65.5**	**87.5**	**92.9**	**75.4**	**93.9**	**97.3**

Table 5: The dense, sparse, and hybrid results of SSR, PTT, and MPP on MSCOCO and Flickr30K in a zero-shot setting.

We present the results of RGP variations in the Table D. In this table, we find that the RGP prompt template seems to be sensitive to LLaVA. With various prompt variations, RGP reranking method performs quite differently. An observation is that an RGP prompt can only achieve improvement on one of the text-to-image and image-to-text retrieval settings. Of course, this situation is suboptimal, and it is better to achieve increases under both retrieval settings at the same time. We believe this is an interesting research point to further optimize the zero-shot cross-modal retrieval results.

E Ethical Statement & Risks

This study complied with all applicable ethical guidelines. Formal ethical approval was not required as the research did not involve human participants or animal experiments.

This study introduces some open-source multimodal large language models on HuggingFace and cross-modal retrieval datasets for evaluation. Considering this, we may violate the protocol to cite these works. It must be declared that we have no intention whatsoever to violate the citation protocol. If anyone finds contents that violate the protocol, the corresponding research will be modified.

As for private information, this study uses images containing person or company objects from datasets. However, these datasets are widely used in cross-modal retrieval and text-based person retrieval tasks. Considering this, it is conceived that this study does not violate privacy laws or contain other offensive information. In addition, because of the use of open-source benchmarks, this study does not attend to any recruitment and payment.

When writing this paper, this study uses AI assistance for polishing, translation, and fixing language bugs. However, we ensure that our research is entirely the product of our own ideas and not generated by AI.

Method		MSCOCO		Flickr30K	
		t2i	i2t	t2i	i2t
LLaVA-Next-LLaMA-8B	Dense	34.6	43.3	60.1	72.4
	MPP Sparse	25.8	29.6	52.5	55.6
	MPP Hybrid	37.3	46.0	64.4	74.1
	MPP+RGP$_1$	47.6	32.3	75.8	55.4
	MPP+RGP$_2$	14.2	49.0	21.9	77.3
	MPP+RGP	14.1	50.7	20.3	79.2
LLaVA-Next-Mistral-7B	Dense	34.3	42.7	60.5	72.3
	MPP Sparse	30.5	39.2	57.0	68.4
	MPP Hybrid	38.1	46.1	65.5	75.4
	MPP+RGP$_1$	48.4	29.3	76.9	53.1
	MPP+RGP$_2$	12.8	49.8	18.1	81.4
	MPP+RGP	12.7	49.4	19.1	80.0

Table 6: The r@1 reranking results of different RGP variations.