Suit the Remedy to the Retriever: Interpretable Query Optimization with Retriever Preference Alignment for Vision-Language Retrieval

GuangHao Meng[1,2*], Jinpeng Wang[3*], Jieming Zhu[4], Letian Zhang[1], Yong Jiang[1,2†], Dan Zhao[2†], Qing Li[2]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Pengcheng Laboratory
[3]Harbin Institute of Technology, Shenzhen
[4]Huawei Noah's Ark Lab
{menggh22, wjp20, zlt23}@mails.tsinghua.edu.cn, jiemingzhu@ieee.org,
jiangy@sz.tsinghua.edu.cn, zhaod01@pcl.ac.cn, liq@pcl.ac.cn

Abstract

Vision-language retrieval (VLR), which uses text or image queries to retrieve corresponding cross-modal content, plays a crucial role in multimedia and computer vision tasks. However, challenging concepts in queries often confuse retrievers, limiting their ability to align concepts with visual content. Existing query optimization methods neglect retrievers' *preferences* (i.e., text descriptions that better match their corresponding visual content), resulting in unadapted to the retriever and leading to suboptimal performance. To address this, we propose the Retriever-Adaptive Query Optimization (RAQO), an interpretable framework that rewrites queries based on retriever-specific *preferences*. Specifically, we first leverages multimodal large language Models (MLLMs) and retrieval's feedback to construct the MLLMs-Driven Preference-Aware Dataset Engine (MPADE), which automatically refine queries offline, capturing the retriever's implicit *preferences*. Then, we introduce a "detect-then-rewrite" chain-of-thought rewriting (ReCoT) strategy equipped with a progressive preference alignment pipeline, including three stages: ambiguity detection fine-tuning, query rewriting fine-tuning, and preference rank optimization. This design enables the rewriter to focus on confusing concepts and produce retriever-adapted, high-quality queries. Extensive VLR benchmark experiments have demonstrated the superiority of RAQO in cross-modal retrieval, as well as its interpretability, generalizability and transferability.

Introduction

Vision-language retrieval (VLR), which uses text/image as queries to retrieve corresponding image/text, has garnered significant attention from both academia and industry (Wang et al. 2022b; Zhao et al. 2023; Wang et al. 2023a, 2024b; Zhang et al. 2025b). Existing methods (Radford et al. 2021; Li et al. 2022a; Yu et al. 2022) mainly focus on how to align text and image modalities within a shared semantic space. By leveraging large-scale pre-trained vision-language data and transformer architectures, these approaches have made remarkable progress in improving retrieval performance. Nevertheless, they still struggle with challenging concepts in queries due to the heterogeneity of multimodal data. As illustrated in Figure 1 (a), the top-1 retrieved image shows the ball already in the dog's mouth, failing to capture the visual features of "anticipate," which leads to inaccurate retrieval. This confusion often hinders the retriever from accurately aligning challenging concepts with corresponding visual features, which limits the performance of VLR. To alleviate this issue, some works (Liu et al. 2024b; Meng et al. 2025) have attempted to utilize a rewriter to optimise queries.

However, without guidance from the retriever's *preferences* (i.e., text descriptions that more accurately match their corresponding visual content), existing VLR query rewriting methods are retriever-agnostic and struggle to adapt effectively to the retriever. Consequently, although the rewritten query elaborates on the concept "anticipate", it still fails to assist the retriever in aligning with the concept of "anticipate", leading to inaccurate retrieval (see Figure 1 (b)).

Ensuring the rewriter generates queries that align with the retriever's preferences in VLR poses a challenge. In natural language processing, exploring the retriever's preferences often relies on labor-intensive rewriting vocabularies (Bhogal, MacFarlane, and Smith 2007; Mandal, Khan, and Kumar 2019) and datasets (Peng et al. 2024), making the process costly and time-consuming. In addition, the retriever's preferences in VLR pose new challenges, as they require consideration of the cross-modal alignment between textual and visual. Multimodal large language models (MLLMs) (Jin et al. 2024; Liang et al. 2025) have demonstrated remarkable capabilities in addressing multimodal tasks (Wei et al. 2024; Gao et al. 2024; Lu et al. 2025; Liu et al. 2025). Employing MLLMs as agents offers a feasible solution for the low-cost acquisition of retriever preferences in VLR.

In this paper, we propose Retriever-Adaptive Query Optimization (RAQO), an interpretable query optimization framework, which leverages MLLMs to offline capture the retriever's preferences, and develops a rewriter to align to retriever's preferences. RAQO can generate high-quality

*Co-first author.
†Dan Zhao and Yong Jiang are corresponding authors.

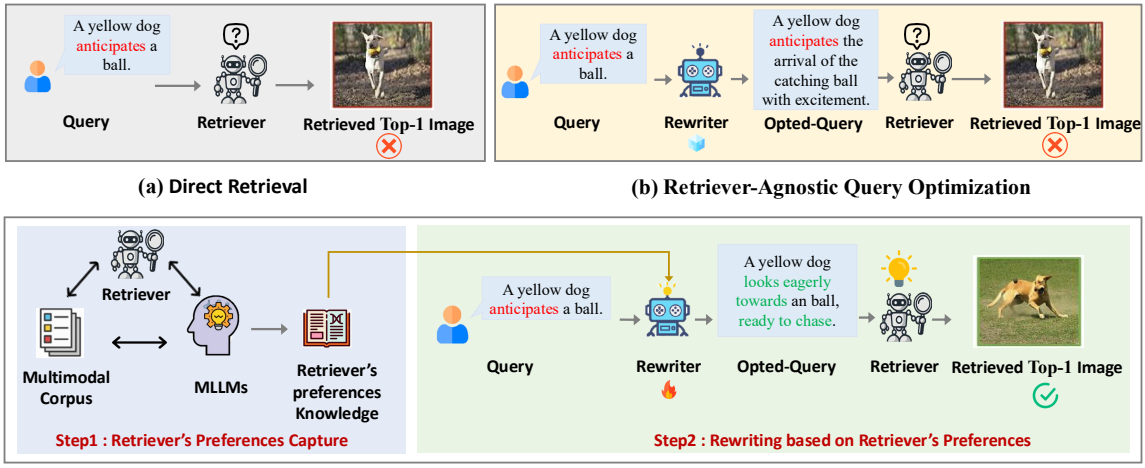

Figure 1: Comparison of Retrieval Paradigms: (a) In direct retrieval methods, the images retrieved are incorrect due to the absence of "anticipate" visual semantics, indicating the retriever cannot align the "anticipate" with the image content. (b) In retriever-agnostic query optimization, the rewriter lacks adaptation to the retriever, thus the retriever still struggles to understand the rewritten query, failing to assist the retriever in aligning with the "anticipated". (c) In contrast, our Retriever-Adaptive Query Optimization, through fine-tuning, learns the retriever's preferences knowledge extracted by MLLMs, adjusting "anticipate" into a visual description that is more aligned to the retriever's preferences, thus ensuring accurate retrieval.

queries that align to retriever's preferences, thereby facilitating more effective cross-modal retrieval. In RAQO, We first construct the MLLMs-Driven Preference-Aware Dataset Engine (MPADE) to offline capture the retriever's preferences. In this engine, MLLMs iteratively adjust the query based on recall feedback from the retriever, automatically identifying the challenging concept of the retriever and generating high-quality query pairs for the rewriting dataset.

Subsequently, we guide the rewriter to perform rewriting based on the retriever's preference. However, enabling the rewriter to understand the rewriting task (Liu and Mozafari 2024) and effectively incorporate these preferences remains challenging. To address this, we design a chain-of-thought rewriting (ReCoT) strategy equipped with a progressive preference alignment pipeline. Specifically, inspired by the human rewriting process, ReCoT decomposes the task into two steps: initially detecting challenging concepts, followed by targeted optimization. To better learn the retriever's preferences for rewriter, we design a novel staged progressive post-training for the rewriter: first, it learns to detect challenging concepts, followed by training on query rewriting tailored to those detected concepts, and ultimately leverages preference rank optimization to capture fine-grained differences across rewriting variants. This progressive strategy enables the rewriter to gradually absorb preference.

In summary, our contributions are as follows:

- We introduce a innovative query optimization for VLR: Retriever-Adaptive Query Optimization (RAQO), which optimizes input queries based on the retriever's preferences, facilitating the retriever to perform more effective alignment between the queries and the visual content.
- We develop an novel MLLMs-Driven Preference-Aware Dataset Engine (MPADE) for offline and automatically capturing retriever's preferences. We employ MLLMs as agents that analyze the retriever's feedback and iteratively optimize queries, generating the rewritten query tailored for vision-language retriever.
- We design a ReCoT equipped with the progressive preference alignment pipeline for better understanding the rewriting and more effectively learns the preference.
- Extensive experiments show RAQO significantly outperforms other advanced query optimization methods in improving VLR performance. Additionally, RAQO is interpretable, generalizable and transferable, performing well across various VLR tasks.

Related Work

Vision-Language Retrieval

VLR aims to to align visual and textual modalities effectively (Zhang et al. 2026). VLR models fall into three categories: single-stream, double-stream, and dual-encoder architectures. Single-stream models (Li et al. 2020; Kim, Son, and Kim 2021) integrate visual and textual inputs into one sequence, using self-attention to enable fine-grained multi-modal interactions for precise alignment. Double-stream models (Li et al. 2021, 2022a; Yang et al. 2022; Tang et al. 2026) separate intra-modal processing from cross-modal fusion, employing co-attention mechanisms that allow for interaction between modalities. Dual-encoder models (Radford et al. 2021; Wang et al. 2022a, 2024a; Li et al. 2025) enhance inference efficiency by projecting visual and textual data into a shared semantic space for similarity assessment, making them suitable for large-scale applications.

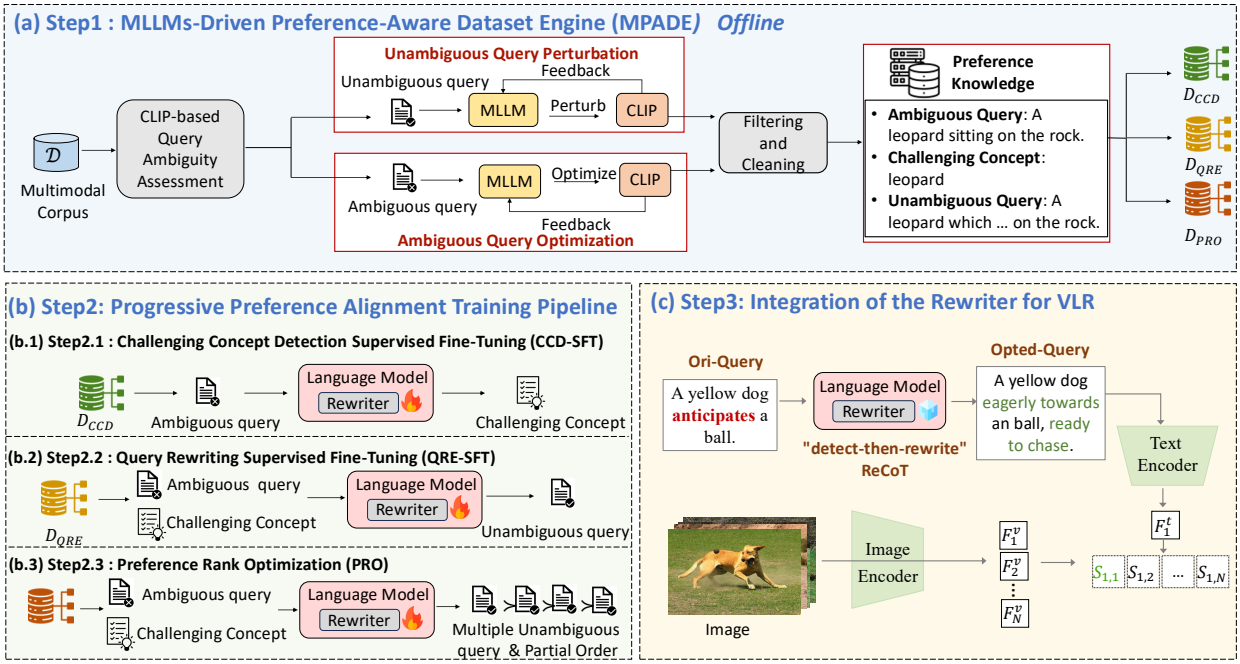

Figure 2: The overview of RAQO. First, we build MPADE to capture the retriever's preferences knowledge, which constructs rewritten query datasets using CLIP's recall feedback and MLLMs' multimodal reasoning. Next, we progressively distill preference into the rewriter through a Progressive Preference Alignment Training Pipeline, comprising CCD-SFT, QRE-SFT, and PRO. Finally, we integrate the rewriter with CLIP to rewrite queries via a "detect-then-rewrite" ReCoT strategy.

Prompt Engineering for VLR

Prompt engineering refines VLR by transforming complex queries into more understandable formats for models. Techniques like Knowledge-CLIP (Pan et al. 2022) leverage external multi-modal knowledge graphs to enhance semantic representations. Similarly, DetCLIP (Yao et al. 2022) use WordNet (Kilgarriff 2000) to augment entity understanding, while RA-CLIP (Xie et al. 2023) uses image retrieval for better cross-modal context. The advent of LLMs has transformed prompt engineering for VLR by tapping into LLM's extensive knowledge (Zeng et al. 2022; Menon and Vondrick 2022; Yu et al. 2024), enhancing the handling of complex queries. DesCLIP (Menon and Vondrick 2022) and CuPL (Pratt et al. 2023) use LLMs to generate visual descriptions and train with enriched queries for better alignment. To reduce noise in descriptions, LaBo (Yang et al. 2023) and Perception-CLIP (An et al. 2023) use scoring functions and learnable weights for selecting optimal descriptions. CLIP-GPT (Maniparambil et al. 2023) integrates a self-attention adapter to filter extraneous information.

Query Rewriting

Query rewriting is a key technique in NLP and e-commerce search, encompassing both discriminative and generative approaches. Discriminative methods (Bhogal, MacFarlane, and Smith 2007; Mandal, Khan, and Kumar 2019; Antonellis, Garcia-Molina, and Chang 2008; Li et al. 2022b) treat it as a retrieval task, expanding semantic scope using predefined vocabularies. Generative methods (Lee, Gao, and Zhang 2018; Qiu et al. 2021), leveraging transformer-based models, generate enriched queries. With the rise of LLMs (Wang, Yang, and Wei 2023; Wang et al. 2023b), these approaches have seen improvements in handling of complex queries. In VLR, query rewriting remains underexplored. Recent works (Liu et al. 2024b; Meng et al. 2026) use LLMs to generate diverse queries via ensemble learning, but these lacked alignment with retriever preferences, limiting effectiveness. NLP methods based on manually crafted vocabularies(Mandal, Khan, and Kumar 2019) and datasets (Peng et al. 2024) address preferences, but they are costly and fail to capture the cross-modal alignment (Li et al. 2024) in VLR.

Methodology

Our proposed query optimizer is designed to automatically rewrite ambiguous user inputs that confuse the retriever, ensuring that the rewritten query retains the original intent while better adaption to the retriever's preferences. As illustrated in Figure 2, RAQO builds upon a VLR backbone and comprises three key steps. First, it constructs an MPADE to capture the retriever's preferences and collect high-quality rewritten queries. Second, the rewriter is fine-tuned through a progressive preference alignment pipeline, comprising Challenging Concept Detection Supervised Fine-Tuning (CCD-SFT), Query Rewriting Supervised Fine-Tuning (QRE-SFT), and Preference Rank Optimization (PRO), to absorb rewriting preferences. Finally, the optimized rewriter is integrated into the retrieval framework and performs rewriting via the ReCoT strategy.

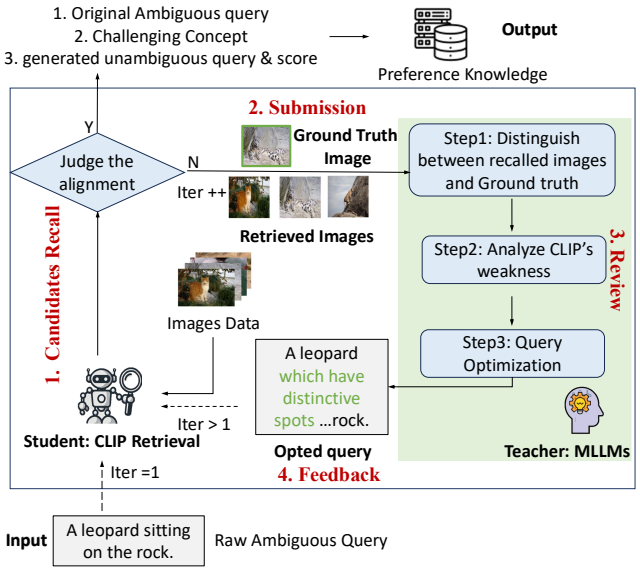

Figure 3: The workflow of our MPADE for optimizing ambiguous queries to capture retriever's preferences.

MPADE

The goal of MPADE is to automatically capture the retriever's preferences by leveraging MLLMs and recall feedback. We define unambiguous queries as those aligned with the retriever's preferences, and ambiguous queries as those that deviate and cause confusion. MPADE generates a large-scale dataset of ambiguous/unambiguous query pairs, where the differences reflect effective query formulations tailored to the retriever's preferences patterns. Additionally, it employs a chain-of-thought process to identify challenging concepts, helping the rewriter better understand retriever preferences. As shown in Figure 2 (a), the workflow consists of four key stages: query ambiguity assessment, ambiguous query optimization, unambiguous query perturbation, and query pair filtering.

Query Ambiguity Assessment We assess query ambiguity using recall feedback from the retriever, with CLIP (ViT-B/32) as the base model. Queries are labeled unambiguous if the ground truth appears in the top-5 retrieval results, indicating alignment with the retriever's preferences. In contrast, ambiguous queries place the ground truth beyond the top-10, reflecting poor interpretability. This classification forms the basis for generating query pairs, allowing MLLMs to explore the retriever's preferences from diverse perspectives.

Ambiguous Query Optimization MLLMs iteratively rewrite ambiguous queries to improve their interpretability for the retriever. As shown in Figure 3, this process involves four steps: candidate recall, submission, review, and feedback. First, candidate images are retrieved using the given query. Next, the retrieved candidate images are submitted along with the ground-truth images for comparison. Third, the differences between the retrieved candidate images and the ground-truth images are reviewed and analyzed using a chain-of-thought process. At this stage, MLLMs identify concepts misunderstood by the retriever and address these by rewriting the query. By analyzing challenging negative samples, MLLMs gain insight into the retriever's weaknesses and adapt the query accordingly. Rewritten queries that meet the unambiguous criteria are added to the unambiguous set. This iterative process yields multiple unambiguous versions per ambiguous query, with text-image similarity used to score each candidate. Additionally, challenging concepts are extracted to guide the ambiguity detector.

Unambiguous Query Perturbation In addition to optimizing ambiguous queries, MLLMs also reverse the process by perturbing unambiguous queries to simulate ambiguity from the retriever's perspective. Using synonym substitution techniques (Mekala, Razeghi, and Singh 2023), core visual concepts are replaced with rare words that preserve meaning but reduce the retriever's ability to match queries with corresponding images. If the perturbed query yields a significantly lower similarity score than the original, it is labeled as ambiguous. Each unambiguous query generates a corresponding ambiguous variant, and the challenging concepts introduced during perturbation are recorded for further use.

Query Pair Filtering To ensure the quality of the dataset, a filtering step is applied after generating ambiguous/unambiguous query pairs. This process includes textual similarity filtering, where the similarity between the generated query and the original query is calculated using a text transformer. Queries with similarity scores falling below a predefined threshold are excluded to ensure that the original user intent is preserved. This filtering step guarantees that the final dataset consists of semantically aligned query pairs, maintaining the integrity of user intent throughout the process.

Dataset Construction and Categorization As shown in Figure 2, the constructed preference knowledge includes a large set of ambiguous/unambiguous query pairs and challenging concepts, covering: original ambiguous queries, their multiple corresponding unambiguous queries with scores; and the pairings of original unambiguous queries with their ambiguous versions. This knowledge is organized into three datasets: (1) D_{CCR} for supervised fine-tuning (SFT) of challenging concept detection, containing ambiguous queries and their associated concepts; (2) D_{QRE} for query rewriting SFT, containing ambiguous queries, challenging concepts, and corresponding unambiguous rewrites; (3) D_{PRO} for preference ranking optimization (PRO), comprising ambiguous queries, their challenging concepts, multiple unambiguous rewrites, and associated scores. Notably, the entire MPADE construction process is fully interpretable.

Query Rewriter in RAQO

We then guide the rewriter to rewrite queries based on the retriever's preference knowledge. To enhance task understanding and absorption of preferences, we propose a chain-of-thought rewriting (ReCoT) strategy with a progressive preference alignment pipeline. ReCoT decomposes rewriting into two steps: detecting challenging concepts and performing targeted optimization. The alignment pipeline comprises

three stages: CCD-SFT, QRE-SFT, and PRO.

Chain-of-Thought Query Rewriting Traditional rewriters may introduce noise or even alter the original semantics. Inspired by the human error-correction process, our method first identifies confusing parts of the query and then performs targeted rewriting. Specifically, the rewriter first identifies concepts in the query that may cause confusion for the retriever, and then refines them while preserving unrelated content. In this way, the rewriter can concentrate on optimizing difficult terms and avoid introducing redundant information, thereby improving the rewriting quality.

Challenging Concept Detection Supervised Fine-Tuning As shown in Fig. 2 (b.1), we perform SFT for the rewriter using the challenge concept detection dataset D_{CCR} built by the preference-aware dataset engine (MPADE). This dataset is specifically designed to train the rewriter to identify challenging concepts that confuse the retriever. Specifically, the challenge concept detection dataset consists of multiple sets of ambiguous queries x paired with their corresponding challenging concepts c. To ensure data balance, we include some unambiguous queries with their corresponding challenging concepts set as *None*. We use $\pi_r(\cdot)$ and θ_r to denote the rewriter to be trained by SFT and its parameters. We train the rewriter by minimizing the cross-entropy loss. The training objective is to optimize the following loss function:

$$L_{\text{CCR}}(\theta_r) = -\mathbb{E}_{(x,c)\sim D_{CCR}} \sum_t \log \pi_r \left(c_t | c_{<t}, x; \theta_r\right). \quad (1)$$

Query Rewriting Supervised Fine-Tuning As shown in Fig. 2 (b.2), QRE-SFT serves as an initial phase that trains a rewriter with basic query rewriting capability through SFT, with the ambiguous/unambiguous query pairs established by the MLLMs-Driven Preference-Aware Dataset Engine. A parallel query corpus, denoted as D_{QRE}, consists of an ambiguous query x and an unambiguous query y. If x corresponds to multiple high-quality queries, we take the highest-quality query as y. We use $\pi_r(\cdot)$ and θ_r to denote the query rewriter to be trained by SFT, which can be any pretrained language model, and its parameters. We train the rewriter by minimizing the cross-entropy loss. The training objective for SFT is to optimize the following loss function:

$$L_{\text{QRE}}(\theta_r) = -\mathbb{E}_{(x,y)\sim D_{QRE}} \sum_t \log \pi_r \left(y_t | y_{<t}, x; \theta_r\right). \quad (2)$$

Preference Rank Optimization To further enhance the rewriter's understanding of the retriever's fine-grained preferences, preference rank optimization, as illustrated in Fig. 2 (b.3), is conducted. This process requires constructing the specialized preference dataset D_{PRO} that we have constructed in MPADE. As described in Section , we generate multiple unambiguous queries for each ambiguous query and obtain text-image similarity scores from the retrieval system, which serve as rewards for preference learning. These scores enable us to rank the queries from high to low based on their alignment with the retriever's preferences.

To enhance fine-grained preference comparisons from a global perspective, we introduce PRO based on the Bradley-Terry model(Song et al. 2024). PRO extends pairwise partial order into general listwise partial order. The PRO loss is expressed by the equation:

$$L_{\text{PRO}}(\theta_r) = -\sum_{j=1}^{k-1} \log \frac{\exp\left(\frac{\pi_{\text{PRO}}(y_j|x;\theta_r)}{\tau_j^{'}}\right)}{\sum_{i=j}^{k} \exp\left(\frac{\pi_{\text{PRO}}(y_i|x;\theta_r)}{\tau_j^{i}}\right)}, \quad (3)$$

where $\tau_j^i = \frac{1}{r(y_j)-r(y_i)}$ reflects the preference gap between candidates, $r(\cdot)$ is the reward defined as the probability of the rewriting generated. The sharper the reward difference, the stronger the supervision. $\tau_j^{'} = \min_{i>j}(\tau_j^i)$ is defined as the minimum temperature among all the candidates to maintain a balance, k denotes the number of candidate unambiguous queries, π_{PRO} and θ_r refer to the policy model and its parameters. L_{PRO} encourages rewriter to assign higher generation probabilities to higher-ranked rewriting.

Experiments

Experiment Setting

Datasets RAQO is evaluated on diverse VLR benchmarks, including Flickr30K (Plummer et al. 2015), MSCOCO (Lin et al. 2014), Flickr30k-CFQ (Liu et al. 2024b), and Llava23K (Liu et al. 2024a). Experiments are conducted on news domain N24News (Wang et al. 2021) and fashion domain Fashion200K (Han et al. 2017) to validate the generalizability. Zero-shot experiments are conducted on WikiDO (Kalyan et al. 2024), Urban1K (Zhang et al. 2025a) and sDCI7K (Urbanek et al. 2024) to validate the transferability.

Baselines We will validate our method on advanced retrieval models, specifically: (1) **CLIP**(Radford et al. 2021), a powerful dual-encoder model pre-trained through contrastive learning; (2) **CoCa**(Yu et al. 2022), a framework that integrates various pre-training paradigms, utilizing its image encoder and unimodal text decoder for retrieval; and (3) **EVA-02-CLIP** (Sun et al. 2023), which employs novel representation learning techniques to enhance CLIP's performance. (4) **BLIP-2** (Li et al. 2023) introduces a framework that connects a frozen image encoder and a pretrained language model through Q-Former. (5) **VLM2Vec** (Jiang et al. 2025) leverages contrastive learning to convert existing MLLMs into embedding-based retrievers.

Our approach will be compared with current query optimization methods, including: (1) **DetCLIP** (Yao et al. 2022): An entity enhancement method that integrates WordNet conceptual knowledge into query entities. (2) **CLIP-GPT** (Maniparambil et al. 2023) An entity description enhancement scheme that incorporates entity visual descriptions generated by LLMs. (3) **RACLIP** (Xie et al. 2023): A retrieval augmentation approach that uses relevant images to enrich the query. (4) **LLMsRewrite** (Liu et al. 2024b): Retriever-agnostic query rewriting based on LLMs.

Implementation details We fine-tune the pre-trained retriever directly, making the process lightweight. Unless otherwise specified, we use the ambiguity detector and Llama2-7B (Touvron et al. 2023) as the query rewriter. During the

Retrievers	Methods	Flickr30K(1K)				MSCOCO(5K)				Flickr30K-CFQ				Llava23K			
		I2T		T2I		I2T		T2I		I2T		T2I		I2T		T2I	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
CLIP	NA	89.1	97.8	74.1	92.6	65.3	85.9	48.1	75.0	73.3	86.8	51.2	75.1	66.3	88.1	61.7	85.1
	+DetCLIP	89.2	97.8	74.6	92.8	65.5	85.9	48.3	75.1	73.9	87.4	52.2	75.6	66.6	88.4	62.1	85.2
	+CLIP-GPT	89.7	98.7	75.2	93.1	66.2	86.2	48.8	75.3	74.0	87.3	52.4	75.6	66.8	88.4	62.1	85.3
	+RACLIP	89.4	98.0	74.5	92.8	65.4	86.1	48.6	75.3	73.8	87.3	51.9	75.4	67.1	88.5	62.2	85.4
	+LLMsRewrite	89.3	98.1	74.5	92.9	65.3	86.0	48.3	75.2	74.2	87.5	52.5	75.7	67.1	88.5	62.0	85.5
	+RAQO	**90.8**	**99.2**	**75.7**	**93.6**	**66.8**	**86.9**	**49.5**	**76.0**	**77.0**	**88.9**	**55.1**	**77.3**	**69.1**	**89.3**	**63.8**	**86.2**
CoCa	NA	85.5	96.5	72.0	91.2	63.9	85.6	45.6	72.1	68.6	84.3	47.8	72.9	64.5	89.0	61.1	85.7
	+DetCLIP	85.6	96.5	72.2	91.2	63.8	85.5	45.8	72.1	69.5	84.8	48.5	73.5	64.6	89.2	61.3	85.7
	+CLIP-GPT	86.2	97.0	72.2	**91.6**	64.3	85.7	46.0	72.2	69.7	84.7	48.6	73.7	64.7	89.3	61.5	85.6
	+RACLIP	85.8	96.6	72.1	91.2	64.1	85.6	45.7	72.1	69.4	84.6	48.3	73.7	64.9	89.4	61.4	86.0
	+LLMsRewrite	86.1	96.7	72.1	91.3	64.2	85.6	45.8	72.1	70.1	85.2	49.0	74.2	65.1	89.6	61.7	86.1
	+RAQO	**86.8**	**97.3**	**72.7**	91.6	**65.0**	**85.9**	**46.6**	**72.7**	**71.4**	**85.8**	**49.9**	**74.7**	**66.2**	**90.3**	**62.7**	**87.0**
EVA-02-CLIP	NA	90.8	98.7	78.9	94.7	69.1	89.2	52.6	78.5	78.1	92.3	57.9	80.4	75.9	93.5	73.7	91.9
	+DetCLIP	90.9	98.6	79.1	94.6	69.3	89.2	52.7	78.5	78.4	92.3	58.1	80.6	76.1	93.4	73.8	91.9
	+CLIP-GPT	91.1	98.7	79.3	94.7	69.4	89.3	52.6	78.6	78.3	92.4	58.0	80.5	75.8	93.5	73.7	92.0
	+RACLIP	90.7	98.6	79.0	94.6	69.1	89.0	52.6	78.5	78.2	92.2	58.1	80.4	76.0	93.6	73.7	91.8
	+LLMsRewrite	91.0	98.6	79.2	94.7	69.2	89.2	52.8	78.6	78.5	92.5	58.2	80.7	76.3	93.9	74.0	92.2
	+RAQO	**91.5**	**98.9**	**79.7**	**95.0**	**69.9**	**89.8**	**53.6**	**79.1**	**79.3**	**92.9**	**58.7**	**81.0**	**76.9**	**94.4**	**74.7**	**92.7**
BLIP2	NA	97.0	100.0	88.7	98.1	83.2	95.9	66.1	86.6	85.2	96.3	66.7	87.4	83.5	97.2	85.7	96.1
	+DetCLIP	97.2	99.9	88.8	98.3	83.3	96.0	66.3	86.8	85.3	96.3	67.1	87.6	83.7	97.2	85.9	96.1
	+CLIP-GPT	97.1	**100.0**	89.0	98.4	83.4	96.1	66.2	86.8	85.5	96.4	66.8	87.5	83.8	97.4	86.0	96.3
	+RACLIP	97.3	100.0	89.1	98.3	83.4	96.0	66.4	86.9	85.2	96.2	66.9	87.6	83.8	97.5	86.0	96.2
	+LLMsRewrite	97.5	99.9	89.1	98.4	83.5	96.0	66.3	86.7	85.4	96.1	67.0	87.7	84.0	97.4	86.1	96.3
	+RAQO	**98.0**	99.9	**89.5**	**98.6**	**84.0**	**96.3**	**66.7**	**87.0**	**86.0**	**96.6**	**67.5**	**88.0**	**84.5**	**97.8**	**86.7**	**96.6**
VLM2Vec	NA	98.1	100.0	89.2	98.7	85.5	96.3	80.2	92.8	87.5	96.8	70.5	89.7	88.6	98.3	90.2	99.1
	+DetCLIP	98.0	100.0	89.4	98.8	85.7	96.3	80.3	93.0	87.6	96.8	70.4	89.7	88.8	98.4	90.3	99.0
	+CLIP-GPT	98.1	100.0	89.2	98.7	85.5	96.3	80.2	92.8	87.8	97.0	70.7	89.9	89.0	98.6	90.4	99.2
	+RACLIP	98.4	100.0	89.4	98.8	85.8	96.5	80.5	**93.1**	88.0	97.1	70.9	90.0	89.1	98.5	90.5	99.1
	+LLMsRewrite	98.3	100.0	89.5	**98.9**	85.9	**96.6**	80.6	93.0	87.9	97.1	71.0	89.8	89.3	98.5	90.4	99.2
	+RAQO	**98.5**	**100.0**	**89.7**	98.8	**86.1**	96.6	**80.7**	92.9	**88.3**	**97.2**	**71.3**	**90.1**	**89.5**	**98.6**	**90.8**	**99.3**

Table 1: Fine-tuning results for image-text retrieval on various VLR benchmark. The visual encoders for CLIP, CoCa, EvA-02-CLIP and BLIP2 are pre-trained ViT-B/32, ViT-B/32, ViT-B/16, and ViT-L respectively. VLM2Vec uses LLaVA-1.6 as backbone. Fair comparison note: All methods are fine-tuned on the same training set of the dataset and our rewriter are only allowed to fine-tune based on preferences extracted from the training set itself, without introducing any external data sources. "NA" denotes setting without any query optimization.

CCD's SFT, we use a learning rate of 2e-5, a batch size of 8, and train for 3 epochs. For the QRE's SFT phase, the learning rate is set to 1e-5, with a batch size of 8, over 3 epochs. In the DPO phase, we use a learning rate of 5e-7, a batch size of 16, for 4 epochs, with a ranking length of 5.

Experiment Result

Excellent performance on Various Retrieval Tasks. As shown in Table 1, we evaluate RAQO on four benchmark datasets and compare it with various query optimization methods using the same training data to ensure fairness. RAQO consistently outperforms baselines across all retrieval tasks. Notably, under the CLIP retriever, it achieves R@1 gains of 3.7%, 3.9%, 2.8%, and 2.1% on I2T and T2I tasks for Flickr30k-CFQ and Llava23K, showing the most significant gains. Compared to entity enhancement methods (e.g., DetCLIP, CLIP-GPT) that focus solely on entity knowledge, and retrieval augmentation methods (e.g., RACLIP) that offer coarse-grained cues, our approach optimizes fine-grained concepts, including entities, interactions, and descriptions. Thus our method displays superior performance in VLR tasks. Additionally, LLMsRewrite underperforms as it fails to adjust rewriting based on retrieve, often producing queries misaligned with retriever preferences.

Significant Improvements Across various Retrievers. Further experiments on various retrievers, as detailed in Table 1, show that applying RAQO significantly boosts performance across all metrics for retrievers such as CoCa, EVA-02-CLIP, and BLIP2. Beyond widely adopted lightweight retrievers, RAQO also achieves the highest improvements on the MLLMs-based retriever VLM2Vec, demonstrating RAQO's effectiveness even for strong retrievers.

Methods	N24News		Fashion200k	
	I2T R@1	T2I R@1	I2T R@1	T2I R@1
CLIP	52.2	51.1	9.5	10.0
+DetCLIP	53.3	51.7	10.0	10.4
+CLIP-GPT	53.5	51.8	9.9	10.4
+RACLIP	53.6	52.0	9.8	10.6
+LLMsRewrite	53.0	51.5	10.1	10.5
+RAQO	**54.7**	**53.0**	**10.9**	**11.5**

Table 2: Fine-tuning results on news domain N24News and fashion domain Fashion200K. Retriever is CLIP (ViT-B/32).

Excellent Generalizability on various Domain. To validate the generalizability of various domains, we evaluate our method on N24News (news) and Fashion200K (fashion). As shown in Table 2, our method improves R@1 for I2T and T2I on N24News and Fashion200K by 2.5%, 1.9%, 1.4%, and 1.5%, significantly outperforming other query optimization methods. We attribute this to MPADE's ability to extract retriever preference knowledge within specific domains, enabling the rewriter to adapt queries with complex, domain-specific terms.

Exceptional Transferability on Unseen Retrieval Tasks. We train RAQO on WikiDO In-Domain (ID) set and conduct zero-shot retrieval on the previously unseen WikiDO

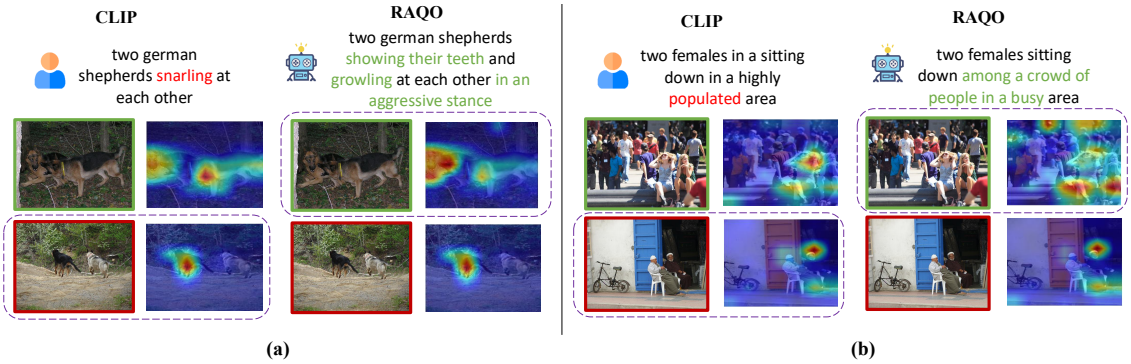

Figure 4: Visualization of text-to-image retrieval with heatmaps for different queries. The image in the dashed box is the recalled result. RAQO rewrites queries into forms more interpretable by the retriever, enabling better alignment with visual content.

Methods	WikiDO ID		WikiDO OOD	
	I2T R@1	T2I R@1	I2T R@1	T2I R@1
CLIP	82.8	81.5	73.4	72.9
+DetCLIP	83.0	81.6	73.5	73.1
+CLIP-GPT	82.9	81.6	73.6	73.0
+RACLIP	83.1	81.7	73.7	73.3
+LLMsRewrite	83.1	81.8	73.6	73.2
+RAQO	**83.7**	**82.4**	**73.9**	**73.6**

Table 3: Comparison of out-of-domain (OOD) performance. Retriever is CLIP (ViT-B/32). The OOD contains images/-queries not seen during fine-tuning on the in-domain (ID).

Rewriter	ReCoT	D_{CCR}	D_{QRE}	D_{PRO}	I2T R@1	T2I R@1
✗	✗	✗	✗	✗	73.3	51.2
✓	✗	✗	✗	✗	74.2	52.5
✓	✗	✓	✓	✓	76.1	54.0
✓	✓	✗	✗	✗	74.4	52.7
✓	✓	✗	✓	✗	75.5	53.7
✓	✓	✓	✓	✗	76.0	54.0
✓	✓	✓	✓	✓	**77.0**	**55.1**

Table 4: Ablation studies on RAQO. The Fine-tuning dataset is Flickr30k-CFQ and retriever is CLIP with ViT-B/32.

Out-of-Domain (OOD) set to evaluate its transferability. As shown in Table 3, RAQO demonstrates strong performance on these previously unseen OOD retrieval tasks. These results indicate RAQO can effectively generalize to unseen retrieval tasks without additional task-specific training. RAQO's transferability in zero-shot settings is attributed to the domain-agnostic retrieval preferences mined by MPADE.

Ablation Study

Effectiveness of Progressive Preference Alignment. As shown in Table 4, incorporating retriever's preference significantly improves VLR performance, regardless of whether ReCoT is used. Specifically, for rewriters equipped with ReCoT, post-training with full preference knowledge improves I2T and T2I R@1 by 2.6% and 2.4%, respectively. To be specific, using D_{QRE} for SFT improves R@1 by 1.1% on I2T and 1.0% on T2I, indicating that D_{QRE} enables the rewriter to acquire rewriting capabilities grounded in preference understanding. Incorporating D_{CCR} for SFT further enhances the rewriter's adaption to retriever preferences, resulting in improved performance. Incorporating preference knowledge D_{PRO} for fine-grained ranking preference optimization further improved R@1 for I2T and T2I by 1.0% and 1.1%. It demonstrates that by learning D_{PRO}, the rewriter can better comprehend the retriever's detailed preferences.

Effectiveness of ReCoT. As shown in Table 4, when incorporating preference knowledge, our "detect-then-rewrite" ReCoT framework further improves R@1 retrieval performance for I2T and T2I by 0.9% and 1.1%, compared to using only the rewriter. This indicates that our ReCoT rewriting mechanism, enabling the rewriter to better understand the rewriting task and generate more precise queries.

Visualization Analysis

To illustrate how RAQO improves query-image alignment, we visualize its effect on the retriever's attention using the Integrated Gradients algorithm (Qi, Khorram, and Li 2019). In Figure 4 (a), CLIP fails to interpret the concept "snarling", resulting in incorrect retrieval. RAQO rewrites it into a more visually grounded phrase "showing their teeth", allowing the retriever to attend to the correct visual cues. Similarly, in Figure 4 (b), CLIP struggles with the term "populated", while RAQO replaces "populated" with a more visually grounded alternative, improving vision-language alignment.

Conclusion

In this paper, we introduce a query optimizer named RAQO for visual-language retrieval, designed to rewrite query concepts that are difficult for retriever to understand into expressions that align with retriever's preferences. Specifically, We first introduce MPADE, which leverages multimodal large language models to capture retriever preferences by analyzing recall performance and iteratively generating high-quality query pairs. We then design ReCoT, equipped with a novel progressive preference alignment pipeline, enabling the rewriter to learn fine-grained reasoning paths for more effective rewriting. Extensive experiments show that RAQO outperforms existing query optimization methods, demonstrating strong generalizability and transferability. Future research could further explore how to integrate user preferences into the rewriter to enrich the application scenarios.

Acknowledgements

This work is supported by the Major Key Project of PCL under grant No. PCL2023A06, the Major Key Project of PCL under grant NO. PCL2025A09, the National Key Research and Development Program of China under grant No. 2022YFB3105000, the National Natural Science Foundation of China under grant No. 624B2088, and the Shenzhen Key Lab of Software Defined Networking under grant No. ZDSYS20140509172959989.

References

An, B.; Zhu, S.; Panaitescu-Liess, M.-A.; Mummadi, C. K.; and Huang, F. 2023. More context, less distraction: Improving zero-shot inference of clip by inferring and describing spurious features. In *ICML Workshops*.

Antonellis, I.; Garcia-Molina, H.; and Chang, C.-C. 2008. Simrank++ query rewriting through link analysis of the click-graph (poster). In *WWW*.

Bhogal, J.; MacFarlane, A.; and Smith, P. 2007. A review of ontology based query expansion. *IPM*.

Gao, D.; Ji, L.; Bai, Z.; Ouyang, M.; Li, P.; Mao, D.; Wu, Q.; Zhang, W.; Wang, P.; Guo, X.; et al. 2024. Assistgui: Task-oriented pc graphical user interface automation. In *CVPR*.

Han, X.; Wu, Z.; Huang, P. X.; Zhang, X.; Zhu, M.; Li, Y.; Zhao, Y.; and Davis, L. S. 2017. Automatic spatially-aware fashion concept discovery. In *ICCV*.

Jiang, Z.; Meng, R.; Yang, X.; Yavuz, S.; Zhou, Y.; and Chen, W. 2025. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In *ICLR*.

Jin, Y.; Li, J.; Liu, Y.; Gu, T.; Wu, K.; Jiang, Z.; He, M.; Zhao, B.; Tan, X.; Gan, Z.; et al. 2024. Efficient multimodal large language models: A survey. *arXiv preprint arXiv:2405.10739*.

Kalyan, T. P.; Pasi, P. S.; Dharod, S. N.; Motiwala, A. A.; Jyothi, P.; Chaudhary, A.; and Srinivasan, K. 2024. Wikido: A new benchmark evaluating cross-modal retrieval for vision-language models. In *ICONIP*.

Kilgarriff, A. 2000. Wordnet: An electronic lexical database.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*.

Lee, M.-C.; Gao, B.; and Zhang, R. 2018. Rare query expansion through generative adversarial networks in search advertising. In *KDD*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*.

Li, S.; Lv, F.; Jin, T.; Li, G.; Zheng, Y.; Zhuang, T.; Liu, Q.; Zeng, X.; Kwok, J.; and Ma, Q. 2022b. Query rewriting in taobao search. In *CIKM*.

Li, T.; Yang, X.; Ke, Y.; Wang, B.; Liu, Y.; and Xu, J. 2024. Alleviating the inconsistency of multimodal data in cross-modal retrieval. In *ICDE*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Li, Y.; Zhen, L.; Sun, Y.; Peng, D.; Peng, X.; and Hu, P. 2025. Deep Evidential Hashing for Trustworthy Cross-Modal Retrieval. In *AAAI*.

Liang, J.; Huang, W.; Wan, G.; Yang, Q.; and Ye, M. 2025. Sculpting lora for harmonizing general and specialized knowledge in multimodal large language models. In *CVPR*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, H.; Gao, K.; Bai, Y.; Li, J.; Shan, J.; Dai, T.; and Xia, S.-T. 2025. Protecting your video content: Disrupting automated video-based llm annotations. In *CVPR*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024a. Visual instruction tuning. *NeurIPS*.

Liu, H.; Song, Y.; Wang, X.; Xiangru, Z.; Li, Z.; Song, W.; and Li, T. 2024b. Flickr30k-cfq: A compact and fragmented query dataset for text-image retrieval. *arXiv preprint arXiv:2403.13317*.

Liu, J.; and Mozafari, B. 2024. Query rewriting via large language models. *arXiv preprint arXiv:2403.09060*.

Lu, Z.; Li, L.; Wang, J.; Feng, Y.; Chen, B.; Chen, K.; and Wang, Y. 2025. CoPRS: Learning Positional Prior from Chain-of-Thought for Reasoning Segmentation. *arXiv preprint arXiv:2510.11173*.

Mandal, A.; Khan, I. K.; and Kumar, P. S. 2019. Query rewriting using automatic synonym extraction for e-commerce search. In *SIGIR Workshops*.

Maniparambil, M.; Vorster, C.; Molloy, D.; Murphy, N.; McGuinness, K.; and O'Connor, N. E. 2023. Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. In *ICCV*.

Mekala, R. R.; Razeghi, Y.; and Singh, S. 2023. Echoprompt: Instructing the model to rephrase queries for improved in-context learning. *arXiv preprint arXiv:2309.10687*.

Meng, G.; He, S.; Wang, J.; Dai, T.; Zhang, L.; Zhu, J.; Li, Q.; Wang, G.; Zhang, R.; and Jiang, Y. 2025. EvdCLIP: Improving Vision-Language Retrieval with Entity Visual Descriptions from Large Language Models. In *AAAI*.

Meng, G.; Wang, J.; Wang, Q.-W.; Ren, X.; and Zhao, D. 2026. Imagine with Layout and Sketch: Enhancing Vision-Language Retrieval with Dual-Stream Multi-Modal Query Refinement. In *AAAI*.

Menon, S.; and Vondrick, C. 2022. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*.

Pan, X.; Ye, T.; Han, D.; Song, S.; and Huang, G. 2022. Contrastive language-image pre-training with knowledge graphs. *arXiv preprint arXiv:2210.08901*.

Peng, W.; Li, G.; Jiang, Y.; Wang, Z.; Ou, D.; Zeng, X.; Xu, D.; Xu, T.; and Chen, E. 2024. Large language model based long-tail query rewriting in taobao search. In *Companion Proceedings of the ACM on Web Conference 2024*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*.

Pratt, S.; Covert, I.; Liu, R.; and Farhadi, A. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*.

Qi, Z.; Khorram, S.; and Li, F. 2019. Visualizing deep networks by optimizing with integrated gradients. In *CVPR Workshops*.

Qiu, Y.; Zhang, K.; Zhang, H.; Wang, S.; Xu, S.; Xiao, Y.; Long, B.; and Yang, W.-Y. 2021. Query rewriting via cycle-consistent translation for e-commerce search. In *ICDE*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Song, F.; Yu, B.; Li, M.; Yu, H.; Huang, F.; Li, Y.; and Wang, H. 2024. Preference ranking optimization for human alignment. In *AAAI*.

Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; and Cao, Y. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Tang, H.; Wang, J.; Zhao, M.; Meng, G.; Luo, R.; and Long Chen, S.-T. X. 2026. Heterogeneous Uncertainty-Guided Composed Image Retrieval with Fine-Grained Probabilistic Learning. In *AAAI*.

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

Urbanek, J.; Bordes, F.; Astolfi, P.; Williamson, M.; Sharma, V.; and Romero-Soriano, A. 2024. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *CVPR*.

Wang, J.; Chen, B.; Liao, D.; Zeng, Z.; Li, G.; Xia, S.-T.; and Xu, J. 2022a. Hybrid contrastive quantization for efficient cross-view video retrieval. In *WWW*.

Wang, J.; Zeng, Z.; Chen, B.; Wang, Y.; Liao, D.; Li, G.; Wang, Y.; and Xia, S.-T. 2022b. Hugs Are Better Than Handshakes: Unsupervised Cross-Modal Transformer Hashing with Multi-granularity Alignment. In *BMVC*.

Wang, J.; Zeng, Z.; Chen, B.; Wang, Y.; Liao, D.; Li, G.; Wang, Y.; and Xia, S.-T. 2024a. Hugs bring double benefits: Unsupervised cross-modal hashing with multi-granularity aligned transformers. *IJCV*.

Wang, J.; Zeng, Z.; Wang, Y.; Wang, Y.; Lu, X.; Li, T.; Yuan, J.; Zhang, R.; Zheng, H.-T.; and Xia, S.-T. 2023a. Missrec: Pre-training and transferring multi-modal interest-aware sequence representation for recommendation. In *MM*.

Wang, L.; Qin, Y.; Sun, Y.; Peng, D.; Peng, X.; and Hu, P. 2024b. Robust contrastive cross-modal hashing with noisy labels. In *MM*.

Wang, L.; Yang, N.; and Wei, F. 2023. Query2doc: Query expansion with large language models. *arXiv preprint arXiv:2303.07678*.

Wang, S.; Scells, H.; Koopman, B.; and Zuccon, G. 2023b. Can chatgpt write a good boolean query for systematic review literature search? In *SIGIR*.

Wang, Z.; Shan, X.; Zhang, X.; and Yang, J. 2021. N24news: A new dataset for multimodal news classification. *arXiv preprint arXiv:2108.13327*.

Wei, Y.; Wang, Z.; Lu, Y.; Xu, C.; Liu, C.; Zhao, H.; Chen, S.; and Wang, Y. 2024. Editable scene simulation for autonomous driving via llm-agent collaboration. In *CVPRW*.

Xie, C.-W.; Sun, S.; Xiong, X.; Zheng, Y.; Zhao, D.; and Zhou, J. 2023. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *CVPR*.

Yang, J.; Duan, J.; Tran, S.; Xu, Y.; Chanda, S.; Chen, L.; Zeng, B.; Chilimbi, T.; and Huang, J. 2022. Vision-language pre-training with triple contrastive learning. In *CVPR*.

Yang, Y.; Panagopoulou, A.; Zhou, S.; Jin, D.; Callison-Burch, C.; and Yatskar, M. 2023. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*.

Yao, L.; Han, J.; Wen, Y.; Liang, X.; Xu, D.; Zhang, W.; Li, Z.; Xu, C.; and Xu, H. 2022. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *arXiv preprint arXiv:2209.09407*.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Yu, Y.; Liang, M.; Yin, M.; Lu, K.; Du, J.; and Xue, Z. 2024. Unsupervised multimodal graph contrastive semantic anchor space dynamic knowledge distillation network for cross-media hash retrieval. In *ICDE*.

Zeng, A.; Liu, X.; Du, Z.; Wang, Z.; Lai, H.; Ding, M.; Yang, Z.; Xu, Y.; Zheng, W.; Xia, X.; et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Zhang, B.; Zhang, P.; Dong, X.; Zang, Y.; and Wang, J. 2025a. Long-clip: Unlocking the long-text capability of clip. In *ECCV*.

Zhang, L.; Meng, G.; Ren, X.; and Wang, J. 2026. Halora: Low-rank Adaptation with Hierarchical Buget Allocation for Efficient Vision-Language Alignment. In *AAAI*.

Zhang, T.; Gao, K.; Bai, J.; Zhang, L. Y.; Yin, X.; Wang, Z.; Ji, S.; and Chen, W. 2025b. Pre-training CLIP against Data Poisoning with Optimal Transport-based Matching and Alignment. In *EMNLP*.

Zhao, M.; Wang, J.; Liao, D.; Wang, Y.; Duan, H.; and Zhou, S. 2023. Keyword-Based Diverse Image Retrieval by Semantics-aware Contrastive Learning and Transformer. In *SIGIR*.