Imagine with Layout and Sketch: Enhancing Vision-Language Retrieval with Dual-Stream Multi-Modal Query Refinement

GuangHao Meng[1,2], Jinpeng Wang[3*], Qian-Wei Wang[1], XuDong Ren[1], Dan Zhao[2*]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Pengcheng Laboratory
[3]Harbin Institute of Technology, Shenzhen
{menggh22, wjp20, wanggw21, rxd21}@mails.tsinghua.edu.cn, zhaod01@pcl.ac.cn

Abstract

Vision-Language Retrieval (VLR) aims to retrieve relevant visual or textual information from multimodal data using language or image queries. However, traditional VLR methods often rely on data-driven shallow semantic alignment and fail to understand the deeper structural and fine-grained entity features of queries, resulting in poor performance on multi-entity layouts and challenging entities. In this paper, we propose the Layout-Aware and Sketch-Enhanced (LASE) VLR framework, which refines query representations by incorporating multimodal layout and sketch knowledge. Specifically, layout knowledge encodes the spatial arrangement of entities, while sketch knowledge refines entity perception by capturing essential structural details. To extract these knowledge representations, we leverage Large Language Models' (LLMs) powerful semantic understanding for layout generation, and Diffusion Models' (DMs) fine-grained cross-modal generative capabilities for sketch generation. However, integrating knowledge into queries may introduce biases and query-specific preferences due to varying visual content and knowledge demands. To address this, we propose the Gated Dual-Stream Knowledge Module (GDKM), which consists of a multi-instance fusion network with a sample-aware gating network. The fusion network aggregates diverse knowledge using multi-head attention to reduce bias, while the gating network adjusts knowledge weights based on query characteristics. Extensive experiments demonstrate that the LASE significantly enhances VLR performance across multiple benchmarks, with superior generalization and transferability.

Introduction

Visual-language retrieval (VLR), which utilizes textual or visual data to retrieve corresponding cross-modal visual or textual information, has become a key focus in multimodal research (Zhao et al. 2023; Wang et al. 2024b; Meng et al. 2025; Zhang et al. 2025b; Meng et al. 2026). However, traditional VLR methods (Radford et al. 2021; Li et al. 2022; Yu et al. 2022) primarily rely on data-driven models for shallow semantic alignment, often overlooking the deep, structured knowledge inherent in text queries. This limitation leads to an insufficient understanding of scene and entity knowledge.

Specifically, VLR involving multi-entity queries faces two primary challenges: **(1) Multi-Entity Layout.** Ambiguous spatial descriptions and limited spatial reasoning hinder the retriever's ability to handle complex layouts. As shown in Figure 1 (a.1), incorrectly retrieved results exhibit significantly lower Spatial Alignment Scores (which quantifies the matching between the query and layout of its corresponding image), indicating a failure to capture spatial structures. For example, in Figure 1 (a.2), the retriever fails to capture the intended spatial relation "about" between "sign" and "bicycle", missing the intended layout. **(2) Cross-Modal Entity Alignment.** The retriever struggles to align textual entities with corresponding visual entities. As shown in Figure 1 (b.1), a noticeable decline in Entity Alignment Score (which quantifies the matching between textual entities and corresponding visual elements) can be observed between correct and incorrect queries. This highlights the retriever's failure to align textual entities with visual features as a key cause of mismatching. For instance, in Figure 1 (b.2), the retriever confuses "bicycle" with "motorcycle", leading to mistakes.

In this paper, we propose a novel Layout-Aware and Sketch-Enhanced (LASE) framework for query refinement, which aims to address the above challenges. Specifically, LASE utilizes the semantic understanding capabilities of Large Language Models (LLMs) (Grattafiori et al. 2024; Lu et al. 2025) and the fine-grained cross-modal generative capabilities of Diffusion Models (DMs) (Betker et al. 2023) to generate *layout* (Lian et al. 2023) and *sketch* (Koley et al. 2024) knowledge relevant to the query, respectively. The term *layout* refers to the spatial arrangement of entities within a query, providing essential spatial cues for modeling inter-entity relationships. As demonstrated in Figure 1 (a.1), incorporating layout significantly enhances Spatial Matching Score, even when the inferred layout deviates from the actual layout of the ground truth image. The term *sketch*, referring to the preliminary visual representations of entities, offers detailed and refined visual features that enhance the retriever's ability to recognize and distinguish similar entities. As shown in Figure 1 (b.1), sketches have proven highly effective in enhancing Entity Alignment Score, demonstrating sketches serve as an ideal medium for conveying entity details.

While integrating layout and sketch provides valuable alignment cues, it also introduces two key challenges: **Knowledge Bias** — Due to the diversity of visual content, queries

*Jinpeng Wang and Dan Zhao are corresponding authors.

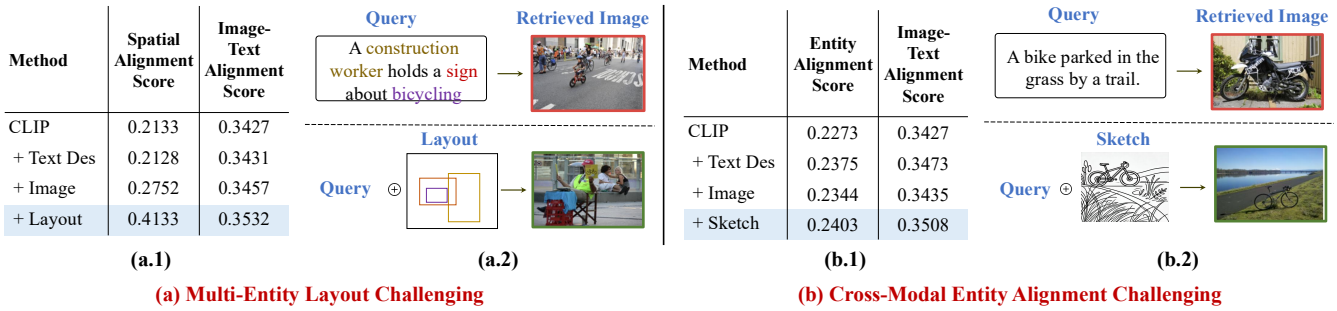

Figure 1: VLR Challenge Examples: Tables (a.1) and (b.1) show a notable drop in Spatial Alignment Score and Entity Alignment Score for incorrectly retrieved queries. (a.2) and (b.2) present examples of incorrect retrievals, where green-bordered images indicate ground truth images, while red-bordered images represent mismatches.

correspond to multiple reasonable layouts and sketch styles. Relying solely on a single knowledge instance can lead to overfitting and thus bring a knowledge bias. **Query-specific Knowledge Preferences** — Different queries require different types of knowledge: layout is crucial for multi-entity queries in Figure 1 (a), while sketch is more effective for queries involving fine-grained entities in Figure 1 (b).

To mitigate these challenges, we propose the **Gated Dual-Stream Knowledge Module (GDKM)**, which comprises a multi-instance fusion network and a sample-aware gating network. The fusion network leverages multi-head attention to integrate diverse layout and sketch instances generated by LLMs and DMs, reducing knowledge bias. The gating network adaptively weights different knowledge sources based on query characteristics, adapting to query-specific knowledge preferences. Furthermore, to deploy in real-world applications, we design the LASE-lite, a lightweight variant that significantly lowers inference cost while preserving the performance benefits.

The contributions of this work are as follows:

- We propose a Layout-Aware and Sketch-Enhanced (LASE) framework, which utilizes LLMs and DMs to generate layouts and sketches. Layout provides essential spatial knowledge, and the sketch enriches visual details. Incorporating both multi-modal knowledge to refine query effectively improves the retriever's ability to handle multi-entity layout and cross-modal entity alignment.
- To address knowledge bias and query-specific knowledge preferences, we design the Gated Dual-Stream Knowledge Module (GDKM). It combines a multi-instance fusion network to capture diverse knowledge and reduce bias, along with a sample-aware gating network that dynamically adjusts weights based on query characteristics.
- Extensive experiments on various VLR benchmarks demonstrate that LASE significantly improves retrieval accuracy, especially in complex layouts queries.

Related Work

Vision-Language Retrieval

The primary goal of VLR is to align visual and textual modalities. Current VLR models fall into three categories: single-stream, double-stream, and dual-encoder architectures. Single-stream models (Chen et al. 2020; 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 to facilitate alignment. Double-stream models (Li et al. 2021, 2022; Yang et al. 2022; Zhang et al. 2024, 2026) separate intra-modal processing from cross-modal fusion, employing co-attention mechanisms to allow interaction between modalities. Dual-encoder models (Li et al. 2021, 2022; Wang et al. 2022b, 2024a; Tang et al. 2025, 2026) enhance inference efficiency by projecting visual and textual data into a shared semantic space for similarity assessment, making them suitable for large-scale applications.

Knowledge-Enhanced VLR Methods

Traditional VLR methods largely rely on data-driven approaches, often neglecting deeper query-specific knowledge. Knowledge-enhanced VLR approaches aim to improve cross-modal understanding through knowledge augmentation, involving internal knowledge extraction or external knowledge integration. Internal extraction methods derive knowledge directly from data. OA-Trans (Wang et al. 2022c) and Structure-CLIP (Huang et al. 2024) use object features for cross-modal learning, while Coder (Wang et al. 2022a) and ViSTA (Cheng et al. 2022) incorporate commonsense and scene text to enhance image-text alignment. External integration methods enrich VLR by incorporating external knowledge sources. Knowledge-CLIP (Pan et al. 2022), EI-CLIP (Ma et al. 2022) and ACP (Fang et al. 2022) leverage multimodal knowledge graphs for concept-level semantics, Other works (Menon and Vondrick 2022; Pratt et al. 2023; Maniparambil et al. 2023; Yang et al. 2023a) use LLMs to generate fine-grained category descriptions or integrate object concepts from WordNet (Yao et al. 2022). Retrieval-enhanced approaches (Xie et al. 2023) retrieve relevant images as cross-modal cues.

However, the existing knowledge-enhanced VLR Methods primarily rely on text descriptions or image-based enhancements, but struggle with spatial layout and fine-grained details. These limitations result in poor performance in tasks involving multi-entity layout alignment and challenging entity alignment. Notably, while some works (Koley et al. 2024) have explored sketch-based VLR, they typically rely on pre-

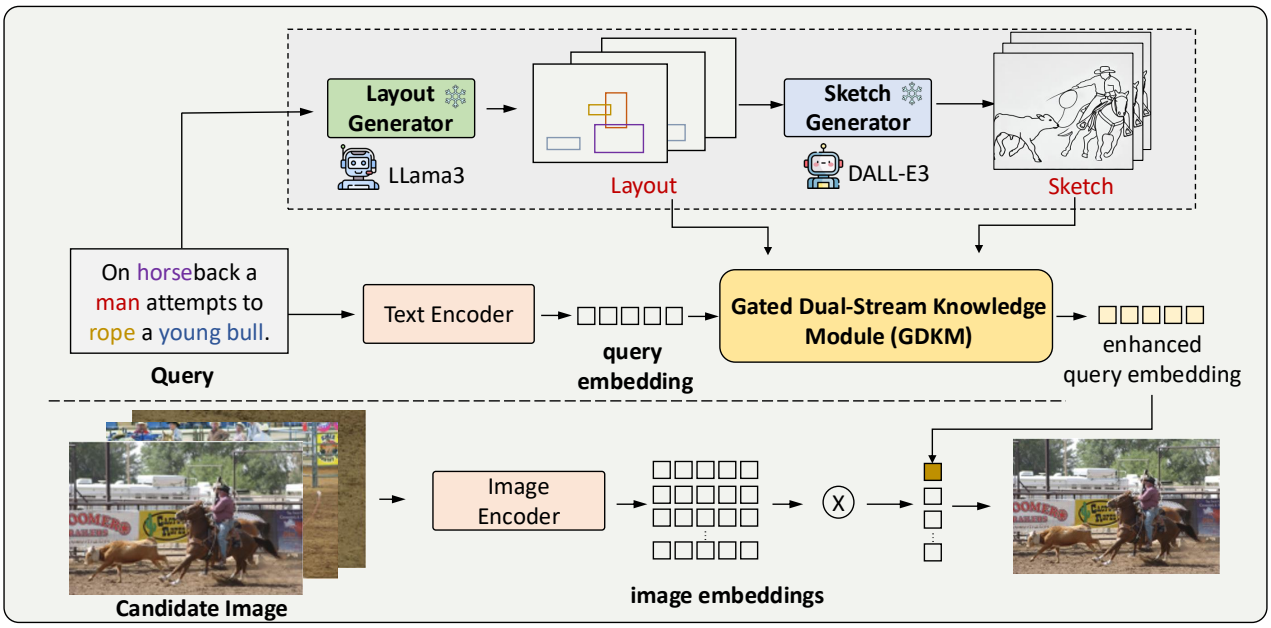

Figure 2: Overview of LASE framework. First, LLMs and DMs are used to infer layout from the query and generate corresponding sketch-based knowledge. Then, GDKM adaptively integrates multi-instance layout and sketch information, producing enriched text embeddings. Finally, the similarity between the enhanced text embeddings and image embeddings is calculated to obtain retrieval results.

constructed sketch libraries and focus primarily on single-entity images in domains like e-commerce. However, many real-world VLR scenarios involve complex multi-entity images without predefined sketch resources. In this paper, we design LASE that incorporates layout generated by LLMs and sketch generated by DMs, providing the models with precise spatial structure and detailed visual cues.

Methods

In this section, we present the details of our proposed LASE framework. Our goal is to enhance VLR models with spatial layout awareness and fine-grained visual details. In our work, we select the simple yet effective dual-encoder model CLIP as the foundational model. Figure 2 shows the overview of the LASE. First, given input queries, we leverage the semantic understanding of LLMS to infer layout information, and the cross-modal generation capability of DMs to generate corresponding sketch cues, respectively. Then the Gated Dual-Stream Knowledge Module (GDKM) is designed to adaptively fuse the multi-instance layout and sketch information to generate enriched text embeddings. Finally, we compute the similarity between the enhanced text embeddings and the image embeddings to obtain the retrieval result.

Knowledge Generation

Layout Generation To accurately represent layout information, we divide it into three components: entity names, bounding box coordinates, and a background description. The bounding box coordinates are represented in the format $(x, y, \text{width}, \text{height})$, where (x, y) represent the center of the bounding box, and width and height specify the bounding box's dimensions. The layout information of a query is expressed as a set: $l = \{(e_1, b_1), \ldots, (e_E, b_E), Z\}$, where E is the number of entities, e_i represents the entity name of the i-th entity, b_i denotes the bounding box coordinate for that entity, and Z is the background description.

LLMs excel in semantic processing and understanding complex textual descriptions, making them ideal for generating coherent and contextually rich layouts. To generate high-quality layout information, we design a prompt template with task instructions and supporting details. Additionally, we provide the LLMs with carefully selected examples based on the task description to ensure the generated layouts adhere to the expected format and structure. The generated layout information can capture each entity's size, quantity, and spatial position, enriching the query with spatial layout details. To reduce potential bias from a single layout, we generate multiple layout variations for each query to provide better diversity and robustness.

Sketch Generation Text descriptions often become lengthy and complex when attempting to convey fine-grained visual details. Images often contain irrelevant background information or distracting features that hinder the transmission of precise details. Sketches provide a concise and efficient medium for fine-grained visual representation. Diffusion models (DMs) (Betker et al. 2023; Li et al. 2025) shows impressive capability to generate high-fidelity, detailed images based on textual input. We leverage DMs to transform the abstract concepts into precise and detailed sketches.

Unlike image classification, VLR often involves multiple

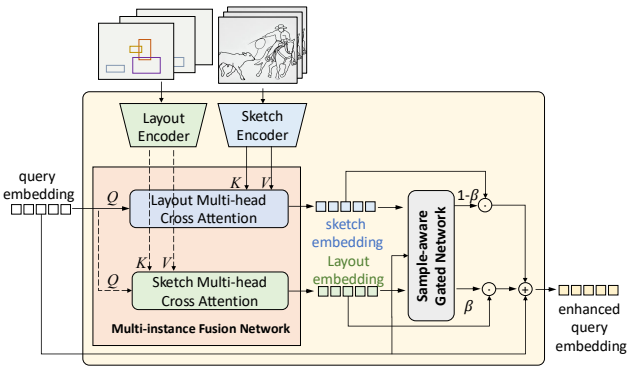

Figure 3: Overview of GDKM. GDKM includes multi-instance fusion network and sample-aware gated network.

entities and their complex spatial relationships. Consequently, directly generating sketches in VLR often results in incomplete or disordered layout information in the sketches. In LASE, we propose a layout-guided strategy for sketch generation. This strategy utilizes a chain-of-thought process, incorporating previously generated layout as prior knowledge for sketch generation, ensuring that the entity arrangements in the sketch better align with the query. In this paper, we use instruction prompts to leverage GPT-4V (Yang et al. 2023b) for invoking DALL-E 3 (Betker et al. 2023) to generate sketches. We design a prompt template that includes a task description and layout information. In addition, we provide high-quality sketch samples that illustrate the desired minimalist style, aiding the model in understanding the expected output. This strategy allows us to generate high-quality, layout-consistent sketches. To accommodate the diverse appearances of entities in images, we generate multiple sketch instances (Zha et al. 2024b, 2025, 2024a), minimizing bias from a single sketch. In our implementation, we generate a corresponding sketch for each layout in section to obtain multiple sketches.

Knowledge Fusion

Gated Dual-Stream Knowledge Module Following the aforementioned process, we have obtained K layouts $\{L_k\}_{k=1}^K$ for each query, and produced the corresponding sketch for each layout, thus obtaining K sketches $\{S_k\}_{k=1}^K$. These layouts and sketches are then input into the Gated Dual-Stream Knowledge Module (GDKM) to generate an enriched query representation. As shown in Figure 3, GDKM initially encodes the layouts and sketches separately using a layout encoder and a sketch encoder, respectively. Subsequently, it effectively integrates multiple layouts and sketches through a multi-instance fusion network. Ultimately, the sample-aware gated network adaptively assigns weights to both layout and sketch knowledge based on query characteristics.

Specifically, the GDKM first extracts the embedding of the layout L_k, denoted as H_k^L, using

$$H_k^L = \phi(L_k), \tag{1}$$

and the embedding of the sketch S_k, denoted as H_k^S, using

$$H_k^S = \psi(S_k), \tag{2}$$

where ϕ and ψ are pretrained unimodal encoders designed specifically for layouts and sketches.

Multi-instance Fusion Network: Inspired by the multi-head attention mechanism's ability to capture diverse instance features across different spaces (Xie et al. 2023), we apply multi-head cross-attention mechanism (Vaswani et al. 2017) to integrate multiple layout and sketch instances of the query, respectively. For the layout instances, we use the query vector H^T as *query*, and use $\{H_k^L\}_{k=1}^K$ as both *keys* and *values* to obtain layout-enhanced embedding:

$$H^L = \text{MultiheadAttn}(H^T, \{H_k^L\}_{k=1}^K, \{H_k^L\}_{k=1}^K). \tag{3}$$

Similarly, for the sketch knowledge, we use the query vector H^T as *query* and use $\{H_k^S\}_{k=1}^K$ as both *keys* and *values* to derive the sketch-enhanced embedding:

$$H^S = \text{MultiheadAttn}(H^T, \{H_k^S\}_{k=1}^K, \{H_k^S\}_{k=1}^K). \tag{4}$$

The multi-instance cross-modal attention mechanism reduces biases that might arise from single-instance reliance, increases the diversity of captured knowledge, and enhances robustness to varied layouts and sketch styles.

Sample-aware Gated Network: Given the diversity of queries, different queries may benefit from sketch and layout knowledge to varying degrees. To address this, we design a Sample-aware Gated Network that adaptively assigns weights to each knowledge type according to the query.

We concatenate the layout-enhanced embedding H^L, the sketch-enhanced embedding H^S, and the query embedding H^T to passed through a multi-layer perception (MLP) (Popescu et al. 2009) to learn the optimal weight distribution. This process is formalized as follows:

$$\beta = \sigma(\text{MLP}(\text{concat}(H^T, H^L, H^S))), \tag{5}$$

where β represents the weight assigned to the layout knowledge, and σ is the sigmoid activation function to ensure β lies between 0 and 1. The weight assigned to the sketch knowledge is set as $1-\beta$. The final enhanced query representation $H_{enhanced}^T$ is calculated as the weighted sum of the query, layout, and sketch representations:

$$H_{enhanced}^T = \beta \cdot H^L + (1-\beta) \cdot H^S + H^T \tag{6}$$

The Sample-aware Gated Network can adaptively allocate attention based on the properties of query, identifying the optimal contribution from layout and sketch to maximize information gain.

Progressive Contrastive Loss In the LASE framework, we introduce a Progressive Contrastive Loss (PCL) aimed at guiding the alignment process by gradually incorporating extra knowledge to improve the model's adaptability to multimodal features. Specifically, PCL initially focuses on basic query and image alignment L_1, gradually to single-stream knowledge (layout or sketch) integration through L_2, and ultimately achieves more complex dual-stream knowledge (layout and sketch) fusion by L_3. Through this incremental approach, the model transitions from basic feature alignment to more complex multimodal integration. Notably, our design is flexible, allowing both the enhancement features of

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
	+DesCLIP	89.5	97.9	74.8	92.8	65.7	85.9	48.4	75.2	73.8	87.6	52.3	75.8	66.9	88.5	62.3	85.3
	+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
	LaBo	89.6	98.5	75.3	93.2	66.4	86.3	48.8	75.4	74.1	87.4	52.2	75.5	67.0	88.6	62.2	85.4
	+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
	+LASE	**90.9**	**99.3**	**75.8**	**93.7**	**66.8**	**86.9**	**49.6**	**76.2**	**76.6**	**88.7**	**54.5**	**77.0**	**68.8**	**89.1**	**63.5**	**86.0**
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
	+DesCLIP	85.8	96.7	72.5	91.4	63.9	85.7	46.1	72.3	69.7	84.8	48.7	73.4	64.8	89.4	61.5	85.8
	+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
	+LaBo	86.3	97.1	72.2	**91.7**	64.3	85.6	46.2	72.4	69.8	84.6	48.6	73.6	64.6	89.3	61.6	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
	+LASE	**86.8**	**97.3**	**72.7**	91.6	**65.0**	**85.9**	**46.6**	**72.7**	**71.2**	**85.7**	**49.5**	**74.5**	**66.1**	**90.2**	**62.5**	**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
	+DesCLIP	90.9	98.7	79.2	94.7	69.4	89.2	52.8	78.6	78.3	92.3	58.3	80.6	76.3	93.6	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
	+LaBo	91.2	98.8	79.2	94.7	69.5	89.3	52.5	78.5	78.4	92.5	58.1	80.5	75.8	93.7	73.8	92.1
	+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
	+LASE	**91.5**	**98.9**	**79.7**	**95.0**	**69.9**	**89.8**	**53.4**	**79.1**	**79.2**	**92.8**	**58.7**	**81.0**	**76.7**	**94.4**	**74.6**	**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
	+DesCLIP	97.3	100.0	88.9	98.4	83.5	**96.2**	66.1	86.5	85.4	96.2	66.9	87.6	83.9	97.4	85.9	96.2
	+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
	+LaBo	97.3	99.9	88.9	97.9	83.5	96.1	66.2	86.7	85.4	96.3	66.8	87.7	83.9	97.5	86.1	96.2
	+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
	+LASE	**97.7**	99.9	**89.6**	**98.4**	**83.7**	96.1	**66.8**	**87.0**	**86.3**	**96.7**	**67.3**	**88.0**	**84.6**	**97.6**	**86.4**	**96.4**
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
	+DesCLIP	98.1	100.0	89.3	98.9	85.6	96.4	80.4	93.0	87.7	96.8	70.6	89.8	88.8	98.3	90.5	99.2
	+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
	+LaBo	98.2	100.0	89.4	98.9	85.8	96.4	80.3	93.0	88.1	**97.2**	70.8	90.0	89.1	**98.7**	90.4	99.0
	+RACLIP	98.4	100.0	89.4	98.8	85.8	96.5	80.4	93.1	88.0	97.1	70.9	90.0	89.1	98.5	90.5	99.1
	+LASE	**98.6**	**100.0**	**89.8**	**99.0**	**86.3**	**96.7**	**80.6**	**93.2**	**88.5**	97.1	**71.5**	**90.2**	**89.3**	98.6	**90.9**	**99.4**

Table 1: Fine-tuning results for image-text retrieval on test set of VLR benchmarks. The visual encoders for CLIP, CoCa, EvA-02-CLIP and BLIP2 are ViT-B/32, ViT-B/32, ViT-B/16, and ViT-L respectively. VLM2Vec uses LLaVA-1.6 as backbone. "NA" denotes setting without any query optimization.

layout and sketches to be optional, which enables diverse configurations in contrastive learning.

We assign the execution ratios P_1, P_2, P_3 of L_1, L_2, L_3 as $\alpha_1 \cdot \left(1 - \frac{t}{T}\right)$, α_2 and $1 - \alpha_1 \cdot \left(1 - \frac{t}{T}\right) - \alpha_2$, respectively, where α_1 and α_2 are tunable parameters, T refers to the number of training epochs, and t denotes the current epoch. The PCL is defined as follows:

$$L_{\text{PCL}} = \begin{cases} L_1, & P_1 = \alpha_1 \cdot \left(1 - \frac{t}{T}\right) \\ L_2, & P_2 = \alpha_2 \\ L_3, & P_3 = 1 - \alpha_1 \cdot \left(1 - \frac{t}{T}\right) - \alpha_2 \end{cases} \tag{7}$$

Experiments

Experiment Setting

Datasets Our method is evaluated on various VLR benchmark datasets. We conduct main experiments on 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.

Baseline We validate LASE on advanced dual-encoder retrieval models, including: 1) **CLIP** (Radford et al. 2021), a powerful dual-encoder model pre-trained via contrastive learning; 2) **CoCa** (Yu et al. 2022), a framework integrating multiple pre-training paradigms, which uses an image encoder paired with a unimodal text decoder for retrieval; and 3) **EVA-02-CLIP** (Sun et al. 2023), which leverages novel representation learning techniques to further improve CLIP's retrieval 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. Additionally, we compare LASE with advanced query optimization methods: 1) **DetCLIP**(Yao et al. 2022) generates object concepts via WordNet; 2) **DesCLIP** (Menon and Vondrick 2022) uses LLMs to generate descriptions and inputs them into CLIP in parallel; 3) **CLIP-GPT** (Maniparambil et al. 2023) creates visual descriptions with LLMs and denoising with a self-attention adapter; 4) **LaBo** (Yang et al. 2023a) selects descriptions with designed functions and a learnable weighting matrix; and 5) **RACLIP** (Xie et al. 2023), a retrieval augmentation technique that enhances queries by integrating relevant images.

Implemented Details LASE is fine-tuned on pre-trained CLIP without pretraining, making the process lightweight. We use Llama3-7B (Grattafiori et al. 2024) and DALL-E 3 (Betker et al. 2023) to generate layout and sketch. Each query is enriched with 4 layout and sketch instances ($K = 4$). For GDKM module, we use 6 layers of cross-attention blocks.

Methods	N24News				Fashion200k			
	I2T		T2I		I2T		T2I	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
CLIP	52.2	72.8	51.1	72.8	9.5	30.7	10.0	30.4
+DetCLIP	53.3	73.4	51.7	73.5	10.0	31.1	10.4	30.7
+DesCLIP	53.5	73.5	51.6	73.6	10.1	31.2	10.5	30.6
+CLIP-GPT	53.5	73.3	51.8	73.4	9.9	31.0	10.4	30.8
+LaBo	53.6	73.5	51.9	73.4	10.0	31.1	10.5	30.7
+RACLIP	53.6	73.6	52.0	73.5	9.8	30.9	10.6	30.5
+LASE	**54.5**	**74.0**	**52.8**	**74.1**	**10.7**	**31.7**	**11.3**	**31.3**

Table 2: Fine-tuning results on news domain N24News and fashion domain Fashion200K. The retriever is CLIP.

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
+DesCLIP	83.1	81.7	73.5	73.0
+CLIP-GPT	82.9	81.6	73.6	73.0
+LaBo	83.0	81.6	73.7	73.1
+RACLIP	83.1	81.7	73.7	73.3
+LASE	**83.6**	**82.3**	**74.1**	**73.7**

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

Method	Component				I2T		T2I	
	Layout	Sketch	Gated Network	PCL	R@1	R@5	R@1	R@5
LASE	✗	✗	–	–	73.3	86.8	51.2	75.1
	✓	✗	–	–	75.0	87.8	52.8	76.0
	✗	✓	–	–	74.7	87.7	52.6	75.8
	✓	✓	✗	✗	75.7	88.1	53.5	76.5
	✓	✓	✓	✗	76.3	88.5	54.2	76.9
	✓	✓	✓	✓	76.6	88.7	54.5	77.0

Table 4: Ablation Studies: Vision Encoder (ViT-B/32), Fine-tuning on Flickr30K-CFQ. ✓ indicates the feature is enabled, ✗ indicates it is disabled, and – denotes configurations not applicable.

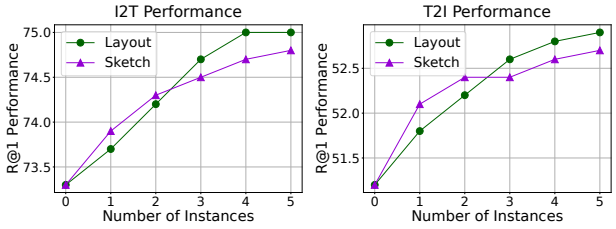

Figure 4: Impacts of the number of instances for layout and sketch. The Fine-tuning dataset is Flickr30K-CFQ.

Main Results

Table 1 presents the fine-tuning results on the Flickr30K, MSCOCO, Flickr30K-CFQ and Llava23K. We compare LASE with several existing query optimization approaches.

Excellent performance on Various Retrieval Tasks. As shown in Table 1, we evaluate various query optimization methods. The results show that LASE significantly outperforms existing methods across various retrieval tasks, consistently improving retriever performance. Notably, under the CLIP framework, LASE achieves R@1 gains of 3.3%, 3.3%, 2.5%, and 1.8% on I2T and T2I for Flickr30k-CFQ and Llava23K. On one hand, compared to entity enhancement methods such as DesCLIP, CLIP-GPT, LaBo and DetCLIP, LASE more effectively captures the essential visual details and spatial cues for text-image alignment. On the other hand, compared to image enhancement RACLIP which relies on global semantics, LASE focuses on spatial relationships and the detailed features of key entities, minimizing unnecessary semantic interference and thereby improving performance.

Significant Improvements Across various Retrievers. Table 1 demonstrates that LASE consistently enhances performance across various retrievers, including CoCa, EVA-02-CLIP, BLIP2, as well as the MLLM-based retriever VLM2Vec, highlighting the effectiveness and general applicability of layout and sketch information as auxiliary features.

Generalizability and Transferability

Excellent Generalizability on various Domain. To verify the versatility of LASE across different domains, we select N24News on news domain and Fashion200K on fashion domain for evaluation. As shown in Table 2, compared to other query optimization methods, our method significantly improves the R@1 for I2T and T2I retrievals on the N24News and Fashion200K datasets, with an increase of 2.3% and 1.7% on N24News, and 1.2% and 1.3% on Fashion200K. In the news domain, the incorporation of layout information is crucial for understanding news content that contains complex visual structures. In the fashion domain, the sketch information effectively captures fashion elements that emphasize visual details. These results not only highlight the flexibility and adaptability of our method but also demonstrate its applicability to a wide range of retrieval tasks.

Exceptional Transferability on Unseen Retrieval Tasks. We train LASE on WikiDO In-Domain (ID) set and conduct zero-shot retrieval on the previously unseen WikiDO Out-of-Domain (OOD) set to evaluate its transferability. As shown in Table 3, LASE exhibits impressive performance on unseen retrieval tasks. This improvement indicates that LASE can adapt to unseen retrieval tasks without the need for additional task-specific training. It highlights the strong transferability of the knowledge enhancement in LASE.

Ablation Studies

Integration of Different Knowledge: Table 4 compares the effectiveness of different combinations of layout and sketch knowledge fusion: using both, only layout, only sketch, and neither. The results indicate that when only layout is used, the R@1 scores for I2T and T2I increased by 1.7% and 1.6%, respectively. When only sketches are used, the R@1 for I2T and T2I increased by 1.4% and 1.4%, respectively. Furthermore, when both sketch and layout information are integrated, the performance improvement is even more significant, with R@1 for I2T and T2I increasing by 2.4% and 2.3%, demonstrating that these modalities complement each other in fine-grained detail and global spatial awareness, thereby enhancing model alignment in complex visual scenarios.

Sample-aware Gated Network: As shown in Table 4, incorporating the samlpe-aware gated network, the R@1 scores

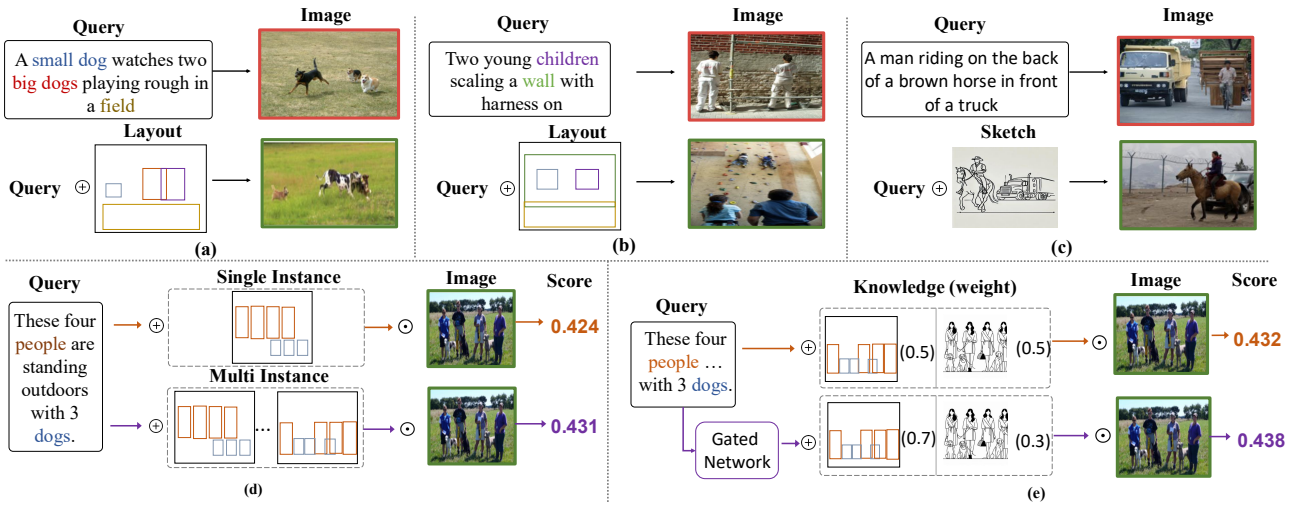

Figure 5: Case study Demonstrating LASE's Superiority: Green borders indicate Ground Truth, Red borders show mismatches.

for I2T and T2I increased by 0.6% and 0.7%, respectively. The gated network adaptively allocates weights to layout and sketch information based on the properties of queries, thereby selecting the most relevant information for the query. Table 5 (e) presents qualitative results of the weight allocation in the gated network. As can be seen, for queries involving multiple entities and complex layouts, the network tends to assign greater weights to layout information. This sample-aware weighting mechanism ensures the flexible and adaptive fusion of knowledge, thereby enhancing retrieval accuracy.

Progressive Contrastive Loss: The progressive learning approach allows the model to gradually refine its understanding of multi-modal information. From Table 4, under the guidance of PCL, the R@1 performance for I2T and T2I respectively improved by 0.3% and 0.3%. The experimental results confirm the effectiveness of PCL in multi-modal retrieval tasks, enabling the model to learn complex cross-modal alignments more efficiently and stably.

The Number of Instances: Figure 4 shows the impacts of the number of instances. The results indicate a noticeable increase in performance improvement with a growing number of instances. This is attributed to reduced biases from single-instance reliance and the addition of diverse auxiliary information, which enhances the method's robustness. Performance gains plateau when the instance count K exceeds 4. Moreover, the impact of multiple layout instances is more significant than that of sketch. As shown in Figure 5 (d), layouts are more prone to knowledge bias, which subsequently impacts VLR performance.

Qualitative Study

Furthermore, we present additional cases to illustrate how layout and sketch information enhance VLR performance. **Multi-Entity Spatial Awareness.** As shown in Figure 5 (a) and (b), layout information helps the model better interpret spatial relationships in multi-entity queries. It captures entity arrangement and relative size (e.g., larger vs. smaller dogs), reducing confusion with images lacking similar compositions and improving retrieval accuracy. **Fine-Grained Entity Differentiation.** Sketches enhance recognition of subtle visual features, aiding in distinguishing similar entities. In Figure 5 (c), the sketch emphasizes key traits (e.g., a man on horseback), enabling CLIP to better differentiate among similar entities. Figures 5 (d) and (e) demonstrate the benefits of multi-instance fusion and the gating mechanism.

Inference Efficiency Limitation

One limitation of LASE its computational efficiency, particularly during inference. While training efficiency can be improved through offline preprocessing of layout and sketch, real-time generation still introduces significant latency in inference. To address this, we develop LASE-lite, a lightweight variant that substantially reduces inference overhead while retaining most of the performance benefits. The design of LASE-lite follows two key principles: (1) replacing the layout generator with a lightweight language model to reduce the cost of spatial reasoning, and (2) replacing sketch generation with a retrieval-augmentation mechanism.

Conclusion

This paper presents LASE, a layout-aware and sketch-enhanced framework for visual-language retrieval (VLR), designed to tackle challenges in multi-entity spatial reasoning and fine-grained entity alignment. By leveraging large language models (LLMs) and diffusion models (DMs), LASE generates query-specific layout and sketch representations, enriching cross-modal alignment with structured spatial and visual cues. To mitigate knowledge fusion bias and adapt to query-specific needs, we introduce the Gated Dual-Stream Knowledge Module (GDKM), which adaptively fuses diverse knowledge sources based on query characteristics. Extensive experiments across multiple VLR benchmarks demonstrate that LASE consistently outperforms existing methods, particularly under complex scene and entity configurations.

Acknowledgments

We sincerely thank the anonymous reviewers and area chairs for their valuable feedback, which has significantly strengthened this manuscript. This work was supported in part by the National Natural Science Foundation of China under grant 624B2088, the Major Key Project of PCL under grant NO. PCL2025A09, and the Shenzhen Key Lab of Software Defined Networking under grant No. ZDSYS2014050917295998.

References

Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Cheng, M.; Sun, Y.; Wang, L.; Zhu, X.; Yao, K.; Chen, J.; Song, G.; Han, J.; Liu, J.; Ding, E.; et al. 2022. Vista: Vision and scene text aggregation for cross-modal retrieval. In *CVPR*.

Fang, S.; Wang, S.; Zhuo, J.; Huang, Q.; Ma, B.; Wei, X.; and Wei, X. 2022. Concept propagation via attentional knowledge graph reasoning for video-text retrieval. In *MM*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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*.

Huang, Y.; Tang, J.; Chen, Z.; Zhang, R.; Zhang, X.; Chen, W.; Zhao, Z.; Zhao, Z.; Lv, T.; Hu, Z.; et al. 2024. Structure-clip: Towards scene graph knowledge to enhance multi-modal structured representations. In *AAAI*.

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*.

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*.

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

Koley, S.; Bhunia, A. K.; Sain, A.; Chowdhury, P. N.; Xiang, T.; and Song, Y.-Z. 2024. You'll never walk alone: A sketch and text duet for fine-grained image retrieval. In *CVPR*.

Li, B.; Wang, J.; Li, Y.; Hu, Z.; Qi, L.; Dong, J.; Wang, R.; Qiu, H.; Qin, Z.; and Zhang, T. 2025. DREAM: Scalable Red Teaming for Text-to-Image Generative Systems via Distribution Modeling. *arXiv preprint arXiv:2507.16329*.

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. 2022. 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, 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*.

Lian, L.; Li, B.; Yala, A.; and Darrell, T. 2023. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*.

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.; 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*.

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*.

Ma, H.; Zhao, H.; Lin, Z.; Kale, A.; Wang, Z.; Yu, T.; Gu, J.; Choudhary, S.; and Xie, X. 2022. Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In *CVPR*.

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*.

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.; Zhu, J.; Zhang, L.; Jiang, Y.; Zhao, D.; and Li, Q. 2026. Suit the Remedy to the Retriever: Interpretable Query Optimization with Retriever Preference Alignment for Vision-Language Retrieval. 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*.

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*.

Popescu, M.-C.; Balas, V. E.; Perescu-Popescu, L.; and Mastorakis, N. 2009. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*.

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*.

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*.

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.; Peng, Y.; Meng, G.; Luo, R.; Chen, B.; Chen, L.; Wang, Y.; and Xia, S.-T. 2025. Modeling uncertainty in composed image retrieval via probabilistic embeddings. In *ACL*.

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*.

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*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NIPS*.

Wang, H.; He, D.; Wu, W.; Xia, B.; Yang, M.; Li, F.; Yu, Y.; Ji, Z.; Ding, E.; and Wang, J. 2022a. Coder: Coupled diversity-sensitive momentum contrastive learning for image-text retrieval. In *ECCV*.

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

Wang, J.; Ge, Y.; Cai, G.; Yan, R.; Lin, X.; Shan, Y.; Qie, X.; and Shou, M. Z. 2022c. Object-aware video-language pre-training for retrieval. In *CVPR*.

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

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

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. 2023a. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*.

Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023b. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*.

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*.

Zha, Y.; Dai, T.; Guo, H.; Wang, Y.; Chen, B.; Chen, K.; and Xia, S.-T. 2024a. Point Cloud Mixture-of-Domain-Experts Model for 3D Self-supervised Learning. *arXiv preprint arXiv:2410.09886*.

Zha, Y.; Li, N.; Wang, Y.; Dai, T.; Guo, H.; Chen, B.; Wang, Z.; Ouyang, Z.; and Xia, S.-T. 2024b. Lcm: Locally constrained compact point cloud model for masked point modeling. *Advances in Neural Information Processing Systems*, 37: 104816–104842.

Zha, Y.; Wang, Y.; Guo, H.; Wang, J.; Dai, T.; Chen, B.; Ouyang, Z.; Yuerong, X.; Chen, K.; and Xia, S.-T. 2025. PMA: Towards Parameter-Efficient Point Cloud Understanding via Point Mamba Adapter. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 16976–16986.

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*.

Zhang, T.; Wang, J.; Guo, H.; Dai, T.; Chen, B.; and Xia, S.-T. 2024. Boostadapter: Improving vision-language test-time adaptation via regional bootstrapping. *NeurIPS*.

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*.