RegRet: Enhancing Region-Level Retrieval in Large Multimodal Models

Xun Liang[1], Honghui Yang[2], Weihang Pan[3], Ruisi Zhao[1], Boyuan Pan[2⋆], Yao Hu[2], Wenxiao Wang[3], Binbin Lin[3⋆⋆], and Deng Cai[1]

[1] State Key Lab of CAD&CG, Zhejiang University
[2] Xiaohongshu Inc.
[3] School of Software Technology, Zhejiang University

Abstract. Region-level retrieval aims to align user-specified image regions with relevant regions or textual descriptions, playing a crucial role in realworld applications such as e-commerce product search and RAG. Although recent Large Multimodal Models (LMMs) have made significant strides in multimodal retrieval, they primarily focus on global-level tasks and struggle to capture effective region-level representations. To bridge this gap, we present ***RegRet***, an LMM-based **Reg**ion-level **Ret**rieval framework that enhances the regional representations without compromising overall global retrieval performance. At its core, RegRet integrates a Region-Aware Encoder to capture detailed regional features while balancing them with the global background context. To further enhance the fine-grained understanding and discriminability of representations, we design a multi-stage training pipeline that includes detailed localized captioning and regional contrastive learning tasks. In addition, considering the absence of region-level contrastive training data and the limited diversity of evaluation tasks in current benchmarks, we introduce the ***REGMB*** benchmark. It comprises 225k contrastive pairs, covering four multimodal retrieval tasks. Extensive experiments validate the effectiveness of our approach. RegRet outperforms strong baselines in the zero-shot setting. Further training with contrastive learning leads to an average improvement of more than 20% on both REGMB and public benchmarks, while achieving comparable or better results on global-level retrieval tasks. The code and data will be released for future research.

Keywords: Large multimodal models · Region-level retrieval ·

1 Introduction

In recent years, multimodal information retrieval has been applied across a variety of domains. In many scenarios, such as e-commerce product search [4, 15, 51], medical image retrieval [22], and retrieval-augmented generation [17], modern industrial retrieval systems typically leverage region-of-interest (ROI) in queries and candidates to enhance region-level matching. For example, as illustrated in Fig. 1,

Fig. 1: The difference between region-level and global-level retrieval. Without the ROI, the negative candidate will be retrieved.

⋆ Corresponding author
⋆⋆ Corresponding author: binbinlin@zju.edu.cn

a shopper may upload a bedroom photo together with a bounding box to find carpets with similar fine-grained texture. Without the ROI, a room that shares the same layout will be assigned a higher similarity score. Although existing studies primarily focus on building image-level representations and assume that ROIs are hard to acquire, a practical question remains underexplored: **How to push the boundaries of retrieval performance when region-level information is available?**

Despite the industrial practical value of the above question, existing multimodal retrieval models [3, 26, 30] are not designed for region-level tasks and thus face two challenges: **(1) Balancing the regional information and background context.** Although some visual understanding LMMs [8,37,49] leverage naive prompting strategies (*e.g.*, crop, ROIAlign) to handle region inputs, directly using them for retrieval will lead to either excessive focus on local regions or the background becoming a distractor for retrieval, leading to suboptimal results as shown in Secs. 5.5.3 and 6. **(2) Comprehensive training and evaluation data.** There are no large-scale region-level contrastive datasets to support training. Meanwhile, existing evaluation benchmarks are also scarce. Most of them are centered on image-to-image task, with limited scenario coverage.

To address the first challenge, we present ***RegRet***, an LMM-based framework for **Reg**ion-level multimodal **Ret**rieval that also supports global-level tasks. At its core, RegRet integrates a Region-Aware Encoder (RAE) alongside the native vision encoder (denoted as context encoder, CE) to build regional embeddings and balance background information. As illustrated in Fig. 2, RAE utilizes cross-attention with global-level features from CE to selectively include necessary background context. The attention modules are organized into a layer-wise paradigm with CE, allowing it to gather information at different semantic levels from all ViT layers. The visual tokens from RAE and CE are aligned with LLM separately, so that it can preserve the image-level retrieval performance after training on region-level datasets. Furthermore, to boost RegRet's retrieval performance and take full advantage of the properties of RAE, we devise a three-stage training pipeline, including (1) RAE pretraining, which employs the Detailed Localized Captioning task [25] under the next-token prediction paradigm to encourage the model to learn detailed regional representations and necessary background context; (2) Pure-text contrastive learning, which use text-only pairs to convert the LMM's language generation capability into embedding capability; and (3) Regional contrastive learning, which use region-level contrastive pairs to further improve the retrieval performance.

To address the second challenge, we establish ***REGMB***, a **Reg**ional **M**ulti-modal retrieval **B**enchmark. It covers four multimodal retrieval tasks [50] and supports ROI-based training and evaluation. REGMB contains 225k region-level contrastive pairs, primarily derived from a manually curated and privacy-sanitized subset from the social media community. The contrastive pairs are first retrieved by SigLip2 [36] and filtered by human annotators. It is further enriched with automatically annotated open-source datasets [1, 2, 14, 18]. Combining the above designs, RegRet outperforms all strong baselines even without additional region-level contrastive training. Specifically, it exceeds the average performance of LMMs by 14.3% on REGMB and 15.0% on existing public benchmarks. When fine-tuned with REGMB, its region-level retrieval performance is further boosted, achieving 7.4% and 7.8% gains over its zero-shot counterpart on the two benchmarks, respectively. Additionally, RegRet delivers on-par or even superior performance to state-of-the-art LMMs on M-BEIR, the global-level retrieval benchmark. Our contributions are as follows:

- We propose RegRet, an LMM that incorporates a Region-Aware Encoder and a three-stage training pipeline, thereby significantly enhancing region-level retrieval capabilities.

– We construct REGMB, the first comprehensive region-level multimodal retrieval benchmark covering four typical scenarios, providing the necessary training and evaluation data that are absent in existing datasets.
– We conduct extensive experiments on REGMB and public benchmarks. RegRet achieves leading region-level performance in both zero-shot and fine-tuned settings, while also preserving global-level retrieval capability.

2 Related Works

LMMs for Multimodal Embedding. Recent advances in LLMs have demonstrated remarkable performance in embedding learning. Early efforts, such as E5 [38] and NV-Embed [21], adapted the generative capabilities of large language models to text retrieval tasks. Building upon this, LamRA [30], Vlm2Vec [13], E5-V [12], and MMEMBED [26] extended the paradigm from text retrieval to multimodal retrieval. More recently, MME5 [3] leveraged synthetic data, while RzenEmbed [11] adopted a refined InfoNCE loss to enhance retrieval performance. Compared to traditional approaches, LMM-based methods exhibit stronger representation and generalization abilities.

Region-Level Representation Learning. A growing body of work focuses on instance-level alignment between image regions and text. RegionCLIP [53], FGCLIP [42], and FineCLIP [16] crop regional features via ROIAlign and align them to fine-grained text descriptions, while LongCLIP [48] and DreamCLIP [52] adopt long captions to achieve finer-grained semantic grounding between image regions and text. However, LMM-based region-level retrieval methods remain underexplored. Some LMMs designed for visual understanding propose some regional prompting strategies, yet these cannot be directly adapted to retrieval tasks. For example, GPT4ROI [49], RegionGPT [8], and GRASP [37] use ROIAlign to crop regional features and concatenate them in the context as auxiliary images. However, naive background cropping may lead to misinterpretation of the region when the ROI is small, while using auxiliary images tends to introduce irrelevant parts of the original image as distractor noise in the final embedding. DAM [25] uses an encoder-decoder-like localized ViT for target-region encoding, but directly adapting it to retrieval tasks leads to conflicts between regional and global features, preventing the model from leveraging data of both types to fully unlock retrieval performance.

Multimodal Information Retrieval Benchmarks. Representative global-level benchmarks include MMEB [13], M-BEIR [40], and MMEB-V2 [31], which cover a wide range of fused-modal tasks (*e.g.*, image-text-to-image retrieval) and extensive training data. In contrast, region-level benchmarks remain limited. A widely adopted evaluation is to reformulate the box classification task of COCO [27] as an image-to-text retrieval problem [41, 53], as shown in Fig. 4. Yet this simplified text-matching paradigm falls short of modern retrieval system requirements. Some early attempts focus on image-to-image retrieval in given scenarios, such as $\mathcal{R}$Oxford [34] for landmark matching and DeepFashion2 [7] for consumer-to-in-shop clothes retrieval. More recently, ILIAS [20] introduces an instance-level benchmark supporting both regional image-image and image-text retrieval. Nevertheless, existing benchmarks are confined to specific cross-modal tasks and do not support training. To date, a comprehensive region-level benchmark that supports training and covers diverse tasks remains absent.

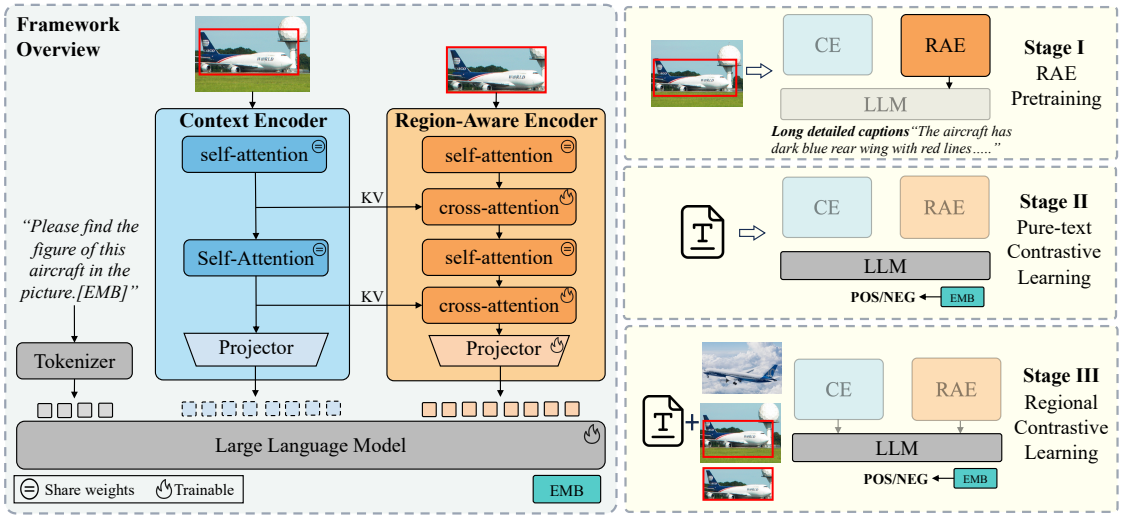

Fig. 2: Framework overview of RegRet (§ 3). The image and its ROI are processed by the native visual encoder (CE) and the RAE, respectively. Each RAE layer applies cross-attention with CE's corresponding layer to extract region-specific signals at matching semantic levels. CE tokens in the dashed box can be used in isolation for global-level retrieval, or jointly used with RAE tokens. An LLM integrates the multimodal tokens and produces the final embedding for retrieval.

3 Methods

In this section, we first formulate the region-level multimodal information retrieval task in Sec. 3.1. Then we elaborate on RegRet in Sec. 3.2, including the way to handle region-level prompts effectively and the architecture of the Region-Aware Encoder. Finally, in Sec. 3.3, we describe the three-stage training strategy designed to boost RegRet's retrieval performance.

3.1 Task Formulation

Region-level multimodal retrieval aims to find the most relevant image or text at the region level. Formally, we define the query set as $Q = \{(q_1, r_{q_1}), (q_2, r_{q_2}), \ldots, (q_M, r_{q_M})\}$, where each query q_k may consist of an image, a text description, or a combination of both. If the query includes an image, it is associated with an ROI r_{q_k}; otherwise, the r_{q_k} is empty. We denote by the operator $q \otimes r_q$ the process of constructing a region-specific query, where the image part of q is cropped to the ROI r_q and the text part remains unchanged. Similarly, we define the candidate set as $C = \{(c_1, r_{c_1}), (c_2, r_{c_2}), \ldots, (c_N, r_{c_N})\}$, where each c_k comprises images, text, or interleaved formats. Given a query q_k, the retrieval process selects the candidate $c_* \in C$ whose regional representation $c_* \otimes r_{c_*}$ is most semantically aligned with $q_k \otimes r_{q_k}$, denoted as:

$$c_* = argmax_{\{c\} \in C} [\langle \Phi(q_k \otimes r_{q_k}),\ \Phi(c \otimes r_c) \rangle], \tag{1}$$

where $\Phi(\cdot)$ denotes the function that embeds the query and candidates into vector representations, and $\langle \cdot, \cdot \rangle$ denotes the similarity function, such as the cosine similarity. In practice, target ROIs can be image patches [45] or region proposals from object detectors [4, 24, 53].

3.2 Model Design

The overall framework of RegRet is shown in Fig. 2. It is built based on an LMM. First, the context encoder processes the entire image query into visual tokens, capturing holistic image content. The region-aware encoder then refines the middle-layer features of the context encoder to extract visual tokens that contain fine-grained information and necessary background context about the ROI. Finally, the LLM backbone integrates these multimodal input tokens into embeddings for retrieval.

3.2.1 Regional Visual Prompts Since LMMs do not support regional inputs, prompting regions to it may largely impact the performance. Though there are several strategies, each of them has inherent flaws for retrieval tasks: (1) **Visual mark.** As shown in Fig. 3(a), it renders marks [43] over the image to indicate the region. This method relies on visual grounding capability, thus sometimes it cannot sufficiently highlight the region, and background information still dominates. (2) **Crop.** Cropping the image region (or visual feature through ROIAlign) may discard the necessary background that helps the model to interpret the region correctly. For example, in Task 1 of Fig. 4(b), the model will mistake the umbrella for white cloth without the background of the shops. (3) **Auxiliary image.** Regions are considered as independent images and concatenated to the tail of the original image as input [37, 46], as depicted in Fig. 3(b). However, the model tends to pay excessive attention to the original image, thus introducing distractor components into the final embedding. A more detailed case study of the above strategies can be found in Sec. 6.

To overcome the shortcomings of the above approaches, we adopt a different strategy: regional inputs are cropped and encoded via a separate Region-Aware Encoder (RAE), as shown in Fig. 3(d). By designing the architecture and training strategy of RAE, we enable it to adaptively balance background context and regional information using learnable parameters, thereby addressing the limitations of naive prompting strategies. The visual tokens from CE are marked in a dashed box, indicating that they are independent of the RAE and can be discarded, leveraged as auxiliary tokens for RAE in regional retrieval, or used in isolation for global-level retrieval.

3.2.2 Region-Aware Encoder We propose the Region-Aware Encoder (RAE), as shown in Fig. 2. The key motivation is to use learnable parameters to adaptively balance the regional and background information. RAE employs a layer-wise coordination paradigm, uses self-attention to refine the regional feature and uses cross-attention to gather background information. Specifically, the image background is first encoded into contextual vision tokens using CE; then, for the ROI, it is first encoded using self-attention, followed by refinement through a cross-attention module. The query comes from the RAE, while the key and value come from the hidden states of the corresponding CE layer. This process can be formulated as:

$$\begin{aligned} h^{i+1}_{\mathrm{RAE}} &= h^{i}_{\mathrm{RAE}} + \alpha \cdot \mathrm{xattn}\big(h^{i}_{\mathrm{RAE}},\, h^{i}_{\mathrm{CE}},\, h^{i}_{\mathrm{CE}}\big)\,, \\ h^{i+1}_{\mathrm{CE}} &= h^{i}_{\mathrm{CE}} + \mathrm{attn}\big(h^{i}_{\mathrm{CE}},\, h^{i}_{\mathrm{CE}},\, h^{i}_{\mathrm{CE}}\big)\,, \end{aligned} \tag{2}$$

where h^{i}_{RAE} is RAE's hidden state on layer i, α is a learnable weight parameter initialized to zero in the beginning, and $xattn$ is an operator that takes arguments in the order (q, k, v). To avoid training a vision encoder from scratch and reduce parameter count, self-attention modules in CE and RAE share weights across corresponding layers. Beyond this, the layer-wise coordination capitalizes on the hierarchical features of ViTs, where shallow layers hold low-level details and deep layers contain high-level semantics [9]. Compared with the encoder-decoder-like methods [25] in Fig. 3(c), which

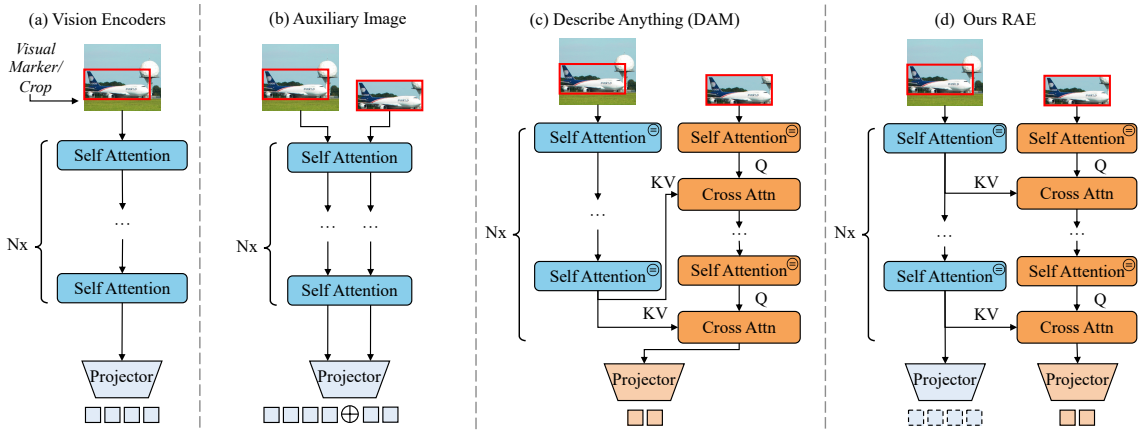

Fig. 3: Comparison of vision backbone architectures and regional prompting strategies in previous methods (§ 3.2). (a) Vanilla ViT with the visual mark or cropping. (b) Auxiliary Image concatenates the target region as an independent image. (c) DAM extracts regional features from the global-level visual embeddings of the last layer. Both ViTs share the same projector. (d) Our RAE adopts separate projectors and layer-wise coordination to balance the regional and global representations.

only use the final-layer hidden states, RAE can capture more detailed cues from early layers, such as material and texture.

During training, since both the RAE and CE output visual tokens, we initially aligned them with the LLM via the same projector. However, we observed that training on region-level data would degrade CE performance, rendering RegRet unable to perform global-level retrieval using CE features. To address this issue, we shifted our perspective: the RAE maintains an independent semantic space, so we use a separate projector to align the RAE with the LLM and freeze the CE parameters. This simple yet effective decoupling design allows us to fully optimize the RAE without compromising the model's original representation capability. As a result, RegRet can retain its global retrieval performance.

3.2.3 LMM For Multimodal Embedding Similar to prior work [10, 30], we add a special token `[EMB]` to the end of the instruction prompt (e.g., `"<image> Summarize the above image and sentence into one word: [EMB]"`). We use the last hidden state after the `[EMB]` as the embedding. The special token functions as a learnable query to summarize the information in the former sentence into an embedding.

3.3 Training Pipeline

Stage-I: RAE Pretraining. This stage serves to teach RAE to balance regional cues with background context and to enhance its fine-grained understanding. We adopt the detailed localized captioning task [25] as the pretext task, compelling the model to combine the background context and generate attribute-rich regional descriptions under the next-token prediction paradigm. We keep the RAE and its connector trainable and freeze the language backbone. Given the token

sequence length T, token x_i, and RAE's trainable parameter θ, the loss can be formulated as:

$$\mathcal{L}_{rae} = -\frac{1}{T}\sum_{t=1}^{T}\log P(x_t \mid x_{<t};\theta). \tag{3}$$

Stage-II: Pure-Text Contrastive Learning. In this stage, we perform contrastive learning with InfoNCE [33] loss to efficiently transfer the LLM's language ability into embedding capabilities. We only use large-scale text-only pairs, as involving image–text pairs brings higher computational cost while yielding comparable performance [30]. After this phase, the model acquires an initial capacity for multimodal retrieval.

Stage-III: Regional Contrastive Learning. In the final stage, we conduct instruction tuning using a mixture of global-level and region-level contrastive pairs. Unlike previous methods [42] which only align image regions with texts, we enable the construction of negative pairs between global and regional samples to enhance the retrieval performance at both levels. The language backbone is optimized with InfoNCE loss, which can be formulated as:

$$\mathcal{L}_{rcl} = -\frac{1}{N}\sum_{i=1}^{N}\log\left[\frac{\exp\Big(\langle \Phi(q_i \otimes r_{q_i}), \Phi(c_i \otimes r_{c_i})\rangle/\tau\Big)}{\sum_{j=1}^{N}\exp\Big(\langle \Phi(q_i \otimes r_{q_i}), \Phi(c_j \otimes r_{c_j})\rangle/\tau\Big)}\right] \tag{4}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product, τ is the temperature parameter, and N is the batch size.

4 REGMB: A Region-Level Multimodal Retrieval Benchmark

In this section, we introduce REGMB[4], a comprehensive **REG**ion-Level **M**ulti-modal Retrieval **B**enchmark. It fills two gaps in current studies: insufficient region-level training data and limited cross-modal evaluation tasks. Some data samples are shown in Fig. 4. REGMB includes a total of 200k contrastive pairs for training and 25k for testing, collected from six diverse datasets and organized into two meta tasks that reflect different levels of retrieval granularity:

Metatask 1: Region-Level Retrieval. This metatask focuses on retrieving similar instances within the ROI. Although background information may aid interpretation, it is unnecessary for successful retrieval. This metatask includes three subtasks covering three retrieval settings: **(i) Task 1:** image-to-text retrieval, **(ii) Task 2:** text-to-image retrieval, and **(iii) Task 3:** image-to-image retrieval. For Tasks 1 and 2, we utilize public datasets with region-level annotations, including COCO [27], SAM [19], and FinHARD [42]. For each region, we use the long, detailed captions generated by DAM [25] and Qwen2.5-VL-72B to increase retrieval difficulty. When building Task 3, a unique challenge arises because *existing public datasets lack cross-scene, region-level pairs of general objects*. For example, in Fashion200k, objects are typically captured at the same angle and under the same lighting conditions in a pure white scene. To overcome this limitation, we construct two data splits, *XGoods* and *XLife*, from a social media platform. We collect a set of image queries, retrieve daily life photos or products, and annotate the relevant pairs by human annotators. As illustrated in Fig. 4(b), the blue shirt appears on the same person but in a different posture and under a different light source. Such variations significantly increase retrieval difficulty. All data are subjected to privacy filtering and manual region-level annotation to ensure high data quality.

[4] More detailed dataset statistics and preprocessing methods are provided in the appendices.

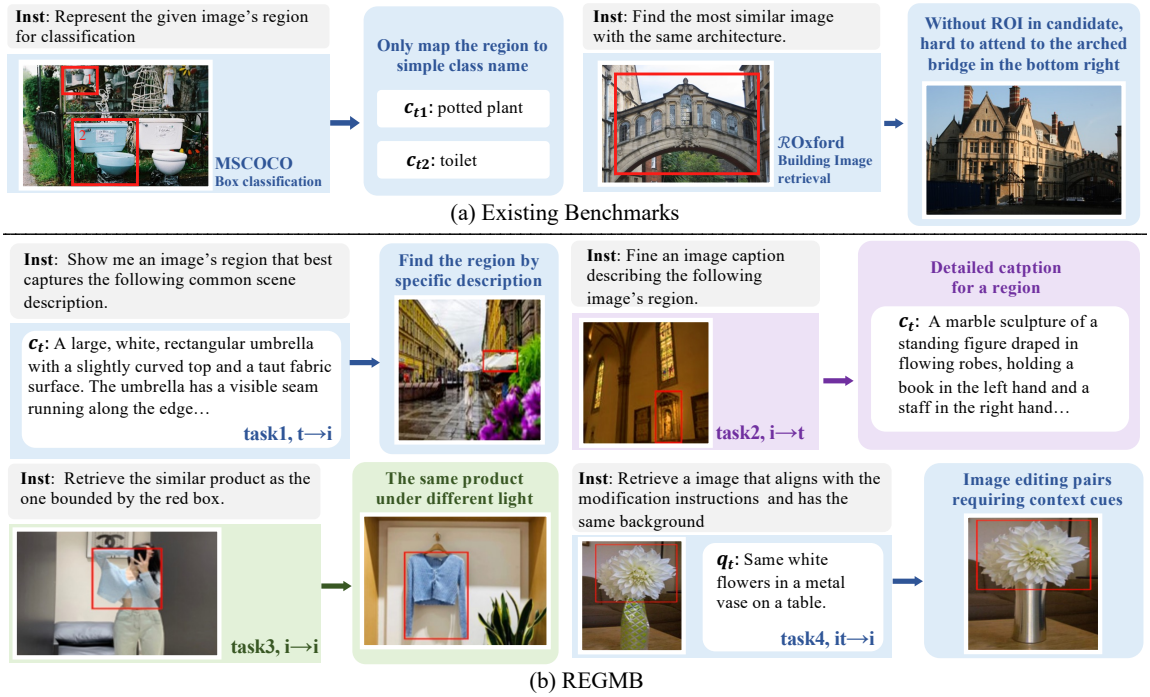

Fig. 4: Examples from REGMB and existing benchmarks (§ 4). REGMB provides complex contrastive pairs and spans a wider range of modalities. *Inst* denotes the task instruction, c_t denotes text candidate, and q_i denotes image query.

Metatask2: Context-Level Retrieval. This setting includes **Task 4**, image-text-to-image retrieval, where background context is critical for accurate retrieval. As illustrated in Fig. 3(b), multiple candidates might contain the flower, but only those with a metal vase in the background are considered correct. We leverage datasets such as VisMin [1] and ImgDiff [14], which naturally include region-level image pairs generated via image editing. Furthermore, we sample a subset from XGoods and automatically annotate it with Qwen2.5-VL-72B as complementary.

5 Experiments

5.1 Implementation Details

We adopt Qwen2-VL [39] models with 3B and 8B parameters as the backbone and train them using the three-stage strategy outlined in Sec. 3.3. In Stage 1, we pretrain the Region-Aware Encoder using 800k region-level image-text pairs from the DAM [25] and PAM [28] datasets. During this stage, we only update the parameters of the RAE's cross-attention layers and its projector. In stage 2, we used the datasets of Natural Language Inference (NLI) [6], HotpotQA [44], and MSMARCO [32], totaling 780k text pairs. In stage 3, we further finetune RegRet on a mixed dataset comprising 1.8M image-level pairs from M-BEIR [40] and 200k region-level pairs from REGMB. More training details are available in the supplementary material.

Table 1: Comparison of methods on REGMB (§ 5.4). RegRet-8B-zs, a zero-shot model without regional contrastive learning, already surpasses most baselines. The regional prompting strategies of baselines are in Sec. 5.2. IT2I denotes image-text-to-image retrieval. † denotes models are trained with REGMB.

Methods	Size	Task 1 (T2I)		Task 2 (I2T)		Task 3 (I2I)		Task 4 (IT2I)			Avg.
		sam	coyo	sam	coyo	xlife	xgoods	vismin	imgdiff	xgoods	
		R@5	R@5	R@5	R@5	R@5	R@5	R@1	R@1	R@5	
CLIP-based											
FG-CLIP [42]	*0.2B*	48.5	58.7	48.9	57.4	50.0	60.7	53.6	**82.2**	66.6	58.5
SigLIP2 [36]	*0.4B*	46.3	74.2	55.9	74.7	76.9	79.0	30.5	51.6	76.7	62.9
DreamLIP [52]	*0.4B*	37.9	63.3	45.9	71.3	49.1	49.3	55.7	54.9	69.6	55.2
FineCLIP [16]	*0.4B*	39.4	72.9	44.2	71.4	60.5	67.7	57.2	28.9	23.7	51.8
EVA-CLIP [35]	*8B*	48.5	58.7	48.9	57.4	50.0	60.7	53.6	**82.2**	66.6	58.5
LMM-based											
MM-EMBED [26]	*8B*	42.5	46.4	35.2	32.1	72.5	80.5	82.9	72.5	87.0	61.3
VLM2VEC [13]	*8B*	39.8	49.9	37.3	48.5	58.7	68.0	79.3	72.2	90.1	60.4
LamRA [30]	*8B*	41.3	55.0	46.7	58.3	75.0	83.3	82.6	62.8	86.0	65.7
RzenEmbed [11]	*8B*	41.9	56.6	59.1	73.6	50.0	68.3	**86.9**	72.5	**98.2**	67.5
mmE5 [3]	*11B*	50.4	55.2	42.2	51.4	75.8	82.0	80.9	75.3	<u>94.4</u>	67.5
Ours											
RegRet-3B†	*3B*	58.1	<u>83.9</u>	70.6	<u>84.3</u>	<u>88.3</u>	<u>91.7</u>	83.1	78.0	81.3	<u>79.9</u>
RegRet-8B-zs	*8B*	<u>59.5</u>	80.3	<u>72.4</u>	80.9	85.1	86.3	81.1	70.8	92.8	78.8
RegRet-8B†	*8B*	**69.1**	**88.9**	**86.5**	**87.3**	**92.3**	**93.6**	<u>83.4</u>	<u>81.1</u>	93.6	**86.2**

5.2 Baselines

We evaluate our model against three categories of baselines: (i) LLM-based methods, such as MME5 [3], VLM2VEC [13], MMEMBED [26], LamRA [30], and RzenEmbed [11]; (ii) CLIP-based methods, including SigLIP2 [36], BLIP2 [23], and UniIR [40]; and (iii) fine-grained region-level models, such as FG-CLIP [42], FineCLIP [29], and DreamLIP [52]. To evaluate the model's zero-shot ability and demonstrate the effectiveness of the architecture design, we further provide **RegRet-8B-zs**, which is trained only on global-level data (*i.e.*, M-BEIR) and not fine-tuned on a region-level dataset. To identify the optimal LMM regional prompting strategy, we investigate multiple candidates. According to the results in Tab. 6, we select the top-performing strategy per task: auxiliary images for I2T and T2I tasks, cropping for I2I task, and visual markers for IT2I task. For FG-CLIP, which supports ROIAlign, we use its native strategy.

5.3 Evaluation Benchmarks

We evaluate RegRet at both regional and global levels. For regional retrieval, we primarily use REGMB, reporting Recall@5 for most tasks and Recall@1 for vismin and imgdiff to avoid metric saturation. To further assess its generalization, we include existing benchmarks like $\mathcal{R}$Oxford Hard split [34], DeepFashion2 [7], and ILIAS [20]. Notably, $\mathcal{R}$Oxford does not have ROIs in its candidate pool. For global-level evaluation, we adopt the M-BEIR [40] benchmark.

Table 2: Generalization evaluation on existing region-level retrieval benchmarks (§ 5.4). The metric for $\mathcal{R}$Oxford-Hard, DeepFashion2, and ILIAS is mAP, recall@1, and mAP@50, respectively. $\mathcal{R}$Oxford only has ROIs in query.

Method	$\mathcal{R}$Oxford-Hard	DeepFashion2	ILIAS		Avg.
	I2I	I2I	I2I	T2I	
VLM2VEC [13]	24.4	6.2	36.1	27.1	23.5
MMEMBED [26]	36.9	12.7	42.7	29.3	30.4
LamRA [30]	42.0	13.2	71.1	62.6	47.2
RzenEmbed [11]	33.9	9.5	42.4	38.6	31.1
mmE5 [3]	36.9	11.1	50.1	39.0	34.2
RegRet-8B-zs	42.8	15.8	75.7	59.1	48.3
RegRet-8B	**45.0**	**20.2**	**86.4**	**72.6**	56.1

Table 3: Performance on the global-level benchmark, M-BEIR (§ 5.4). Although specifically optimized for region-level tasks, RegRet maintains its global-level retrieval capibility and even achieves better results. † denotes models trained with M-BEIR.

Methods	T2I		T2IT		I2T		I2I	IT2T	IT2I		IT2IT		Avg.
	COCO	F200K	EDIS	WebQA	COCO	F200K	NIGHTS	InfoSeek	F200IQ	CIRR	OVEN	InfoSeek	
	R@5	R@10	R@5	R@5	R@5	R@10	R@5	R@5	R@10	R@5	R@5	R@5	
SigLIP [47]	75.7	**36.5**	27.0	43.5	88.2	**34.2**	28.9	25.1	14.4	22.7	41.7	27.4	38.8
BLIP2 [23]	63.8	14.0	26.9	24.5	80.0	14.2	25.4	5.5	4.4	11.8	27.3	15.8	26.1
Qwen2-VL-7B [39]	55.1	5.0	26.2	9.4	46.6	4.0	21.3	22.5	4.3	16.3	43.6	36.2	22.3
UniIR-BLIP$_{FF}$† [40]	79.7	26.1	50.9	79.8	89.9	28.9	**33.0**	22.4	29.2	52.2	55.8	33.0	48.4
UniIR-CLIP$_{SF}$† [40]	81.1	18.0	59.4	78.7	**92.3**	18.3	32.0	27.9	24.4	44.6	67.6	48.9	49.4
LamRA† [30]	81.5	28.7	62.6	81.2	90.6	30.4	32.1	52.1	**33.2**	53.1	76.2	63.3	57.1
RegRet-8B†	**81.6**	28.9	**62.7**	**81.8**	90.3	30.6	31.6	**52.7**	**33.2**	**53.3**	**76.4**	**64.5**	**57.3**

5.4 Analysis

Our main results are shown in Tab. 1 and Tab. 2. For both REGMB and public datasets, our method outperforms strong baselines. Several key observations can be drawn from the results:

(1) The effectiveness of model design. In the zero-shot setting, RegRet-8B-zs outperforms all the baselines on REGMB and public benchmarks. It exceeds the average performance of LMMs that employ regional prompting strategies by 14.3% and 15.0%, respectively. It is even on par with the fine-tuned leading LMM in Tab. 5. For $\mathcal{R}$Oxford, which does not have ROIs in candidates, our method can still perform well. This is mostly due to the RAE's balance between regional and background information.

(2) The effectiveness of training data and strategy. After training on REGMB, the 3B model already surpasses some 11B models (e.g., mmE5), and RegRet-8B achieves an extra average improvement of 7.4% compared to the zero-shot version. The performance gain generalizes to unseen public benchmarks, yielding an average 22.8% improvement, demonstrating the effectiveness of our data and training strategy.

(3) RegRet can preserve or even enhance global-level retrieval performance. As reported in Tab. 3, we observe performance gains compared to LamRA on the majority of splits in M-BEIR. Since novel models like mmE5 and RzenEmbed have made many improvements to training data and

Table 4: Ablations on the ViT architectures shown in Fig. 3 (§ 5.5.1). All models are trained on REGMB and M-BEIR based on the RegRet-8B.

Architecture	xattn	separate proj.	layer-wise	Task1	Task2	Task3	Task4	REGMB Avg.	M-BEIR Avg.
Auxiliary Image	✗	✗	✗	69.8	82.7	92.6	**86.7**	83.4	57.1
RAE-sharep	✓	✗	✗	77.2	83.5	51.7	52.2	66.0	27.2
RAE-encdec	✓	✓	✗	76.6	83.8	89.4	81.3	82.6	57.1
RAE	✓	✓	✓	**79.0**	**86.9**	**92.9**	86.0	**86.2**	**57.3**

Table 5: Ablations on the training data (§ 5.5.2). We also trained LamRA from scratch with its original dataset and REGMB to show the effectiveness of our data.

Model	Size	Language	Image-level	Regional	Task1	Task2	Task3	Task4	Avg.
RegRet	*3B*	✓			61.3	65.4	74.5	71.1	68.4
		✓	✓		65.1	62.4	64.6	80.1	69.4
		✓	✓	✓	**71.0**	**77.4**	**90.0**	**80.8**	**79.9**
LamRA	*8B*	✓	✓	✓	64.0	79.2	89.0	**87.8**	80.9
RegRet	*8B*	✓	✓	✓	**79.0**	**86.9**	**92.9**	86.0	**86.2**

the loss function, we argue that LamRA is the most reasonable baseline for examining global-level performance. The gains are particularly notable on IT2IT tasks like InfoSeek, where it increases by 1.2%. This indicates that training RegRet on regional contrastive data can at least preserve global-level retrieval ability.

(4) Inherent tension exists between local and global representation learning. Models that perform well in Tasks 1–3 (*e.g.*, SigLIP2) may underperform in Task 4, and vice versa (*e.g.*, MME5). For Tasks 1-3, the model needs to exclude background noise, while the background information is crucial in Task 4. Models that focus on region-level features are less effective at global-level understanding. Therefore, RegRet also slightly underperforms some models on Task 4. Despite this, our method achieves a sweet spot in this trade-off by decoupling RAE and CE, enabling it to perform effectively across all tasks.

5.5 Ablations

In this section, we conduct ablations to further verify the effectiveness of each module. We choose LamRA as the LMM for comparison because it shares a similar training procedure and dataset with RegRet, without synthetic data (*e.g.*, mmE5) and with an improved InfoNCE loss (*e.g.*, RzenEmbed). Therefore, it ensures a fair validation of RegRet's performance.

5.5.1 RAE Architecture To validate the effectiveness of each design in RAE, we compare it with different architectures, as shown in Tab. 4. RAE-sharep use a single projector using the last-layer's hidden states. RAE-encdec adds separate projectors, and RAE further employs the layer-wise coordination. All the variants are based on RegRet-8B and undergo the same training process. We can observe that using separate projectors can prevent interference between local and global information. RAE-encdec avoids the giant performance degradation compared to RAE-sharep. Moreover, although the auxiliary image strategy is comparable to the RAE-encdec on REGMB, it still underperforms on Tasks 1 and 2, as analyzed in Sec. 6. RAE incorporates layer-

wise coordination and further enhances region-level performance, as it can align CE and RAE features across different semantic levels.

5.5.2 Training Data. Tab. 5 presents ablations on different training data sources. We compare the use of pure-text data from Stage-II, image-level data from M-BEIR, and regional-level data from REGMB. We observe a notable boost in regional retrieval accuracy after integrating region-level contrastive pairs. Training LamRA with REGMB also improves its performance (67.7%⟶80.9%). This reveals the effectiveness of our data.

5.5.3 Regional Prompt Strategy To demonstrate that directly transferring regional prompting strategies to the retrieval task underperforms RAE, we equip LamRA with multiple strategies and compare it with RegRet-8B-zs. The results are reported in Tab. 6. We evaluate three prompt methods as described in Sec. 3.2, including visual mark, crop, and auxiliary image. Results show that cropping and visual mark are close in performance, while the auxiliary image yields substantial improvements. However, even with the most effective strategy, LamRA still falls short of RegRet, suggesting that simply adding or cropping an image is insufficient to achieve the same performance gains.

Table 6: Zero-shot performance of RegRet and LMMs with different prompting strategies on REGMB (§ 5.5.3). Only using prompts to adapt LMMs models to region-level retrieval proves suboptimal.

Prompt	Task1	Task2	Task3	Task4	Avg.
Crop	55.3	28.3	79.2	59.2	55.9
Visual mark	39.8	40.5	55.7	77.1	55.9
Auxiliary image	48.1	52.5	70.4	75.2	63.1
RegRet-8B-zs	**69.9**	**76.7**	**85.7**	**81.6**	**78.8**

6 Case Study

In this section, we conduct a detailed case study to illustrate why RegRet outperforms existing LMMs across various regional prompting strategies. Retrieval samples from RegRet-8B and LamRA are listed in Fig. 5. We can make the following observations: **(1) RegRet can incorporate background context.** In Fig. 5(c), it selects the image showing a woman in front of a café. Although other candidates wear similar clothing, they are rejected because they do not meet the user's background requirements. **(2) RegRet can discard background distractors, while auxiliary image may introduce it.** As shown in Fig. 5(e), RegRet only focuses on the racecar within the bounding box. However, when using the baseline LMM with auxiliary images, it incorrectly matches the image to "crowd", which dominates the background. **(3) Cropping may lose background for accurately interpreting ROI.** In Fig. 5(b), when LamRA uses cropped regions to retrieve, it mistakenly interprets a flower for the golden fish decoration. If it is fed the green background, the model can understand that the candidate is a handicraft rather than a building decorative element. **(4) RegRet has fine-grained visual understanding ability.** In Fig. 5(a), while the baseline returns leather black jackets, RegRet captures the material feature and finds the correct cloth ones. This ability arises from the RAE pretraining stage using detailed captions.

7 Conclusion

We present RegRet, a novel framework designed for diverse region-level multimodal retrieval tasks. Equipped with the Region-Aware Encoder, it significantly enhances region-level representation. We

Fig. 5: Queries and top three candidates from REGMB (§ 6). All the results are retrieved from the around 5k candidate pool of each subtask. The positive results are labelled with green checkmarks.

further build the REGMB benchmark to enable the training and evaluation of regional retrieval models. Compared to previous approaches, RegRet sets a new state-of-the-art in performance. Future work will explore extending RegRet's ability to visual document retrieval and integrating it into retrieval-augmented generation systems.

Acknowledgements

This work was supported in part by the National Nature Science Foundation of China (Grant No: 62273303), in part by Yongjiang Talent Introduction Programme (2022A-240-G).

References

1. Awal, R., Ahmadi, S., Zhang, L., Agrawal, A.: Vismin: Visual minimal-change understanding (2024)

2. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset (2022)
3. Chen, H., Wang, L., Yang, N., Zhu, Y., Zhao, Z., Wei, F., Dou, Z.: mme5: Improving multimodal multilingual embeddings via high-quality synthetic data. arXiv preprint arXiv:2502.08468 (2025)
4. Dong, X., Zhan, X., Wu, Y., Wei, Y., Kampffmeyer, M.C., Wei, X., Lu, M., Wang, Y., Liang, X.: M5product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21220–21230. IEEE Computer Society (2022)
5. Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., Colombo, P.: Colpali: Efficient document retrieval with vision language models (2025), https://arxiv.org/abs/2407.01449
6. Gao, T., Yao, X., Chen, D.: Simcse: Simple contrastive learning of sentence embeddings. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2021)
7. Ge, Y., Zhang, R., Wu, L., Wang, X., Tang, X., Luo, P.: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. CVPR (2019)
8. Guo, Q., De Mello, S., Yin, H., Byeon, W., Cheung, K.C., Yu, Y., Luo, P., Liu, S.: Regiongpt: Towards region understanding vision language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13796–13806 (2024)
9. He, T., Tan, X., Xia, Y., He, D., Qin, T., Chen, Z., Liu, T.Y.: Layer-wise coordination between encoder and decoder for neural machine translation. Advances in Neural Information Processing Systems **31** (2018)
10. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of the International Conference on Machine Learning (2021)
11. Jian, W., Zhang, Y., Liang, D., Xie, C., He, Y., Leng, D., Yin, Y.: Rzenembed: Towards comprehensive multimodal retrieval. arXiv preprint arXiv:2510.27350 (2025)
12. Jiang, T., Song, M., Zhang, Z., Huang, H., Deng, W., Sun, F., Zhang, Q., Wang, D., Zhuang, F.: E5-v: Universal embeddings with multimodal large language models. arXiv preprint arXiv:2407.12580 (2024)
13. Jiang, Z., Meng, R., Yang, X., Yavuz, S., Zhou, Y., Chen, W.: Vlm2vec: Training vision-language models for massive multimodal embedding tasks. arXiv preprint arXiv:2410.05160 (2024)
14. Jiao, Q., Chen, D., Huang, Y., Ding, B., Li, Y., Shen, Y.: Img-diff: Contrastive data synthesis for multimodal large language models (2024), https://arxiv.org/abs/2408.04594
15. Jin, Y., Li, Y., Yuan, Z., Mu, Y.: Learning instance-level representation for large-scale multi-modal pretraining in e-commerce. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11060–11069 (2023)
16. Jing, D., He, X., Luo, Y., Fei, N., Wei, W., Zhao, H., Lu, Z., et al.: Fineclip: Self-distilled region-based clip for better fine-grained understanding. Advances in Neural Information Processing Systems **37**, 27896–27918 (2024)
17. Kim, J., Tao, R., Sharma, S., Wang, J., Sun, K., Lin, Z., Moon, S., Mathias, L., Kumar, A., Ji, H., Dong, X.L.: Pixel-grounded retrieval for knowledgeable large multimodal models (2026), https://arxiv.org/abs/2601.19060
18. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
19. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
20. Kordopatis-Zilos, G., Stojnić, V., Manko, A., Šuma, P., Ypsilantis, N.A., Efthymiadis, N., Laskar, Z., Matas, J., Chum, O., Tolias, G.: Ilias: Instance-level image retrieval at scale (2025), https://arxiv.org/abs/2502.11748
21. Lee, C., Roy, R., Xu, M., Raiman, J., Shoeybi, M., Catanzaro, B., Ping, W.: Nv-embed: Improved techniques for training llms as generalist embedding models (2024)

22. Lee, H.H., Santamaria-Pang, A., Merkow, J., Oktay, O., Pérez-García, F., Alvarez-Valle, J., Tarapov, I.: Region-based contrastive pretraining for medical image retrieval with anatomic query. arXiv preprint arXiv:2305.05598 (2023)
23. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the International Conference on Machine Learning (2023)
24. Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., Wang, H.: Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 2592–2607 (2021)
25. Lian, L., Ding, Y., Ge, Y., Liu, S., Mao, H., Li, B., Pavone, M., Liu, M.Y., Darrell, T., Yala, A., et al.: Describe anything: Detailed localized image and video captioning. arXiv preprint arXiv:2504.16072 (2025)
26. Lin, S.C., Lee, C., Shoeybi, M., Lin, J., Catanzaro, B., Ping, W.: Mm-embed: Universal multimodal retrieval with multimodal llms (2024), https://arxiv.org/abs/2411.02571
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
28. Lin, W., Wei, X., An, R., Ren, T., Chen, T., Zhang, R., Guo, Z., Zhang, W., Zhang, L., Li, H.: Perceive anything: Recognize, explain, caption, and segment anything in images and videos (2025)
29. Lin, W., Chen, J., Mei, J., Coca, A., Byrne, B.: Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering (2023)
30. Liu, Y., Zhang, Y., Cai, J., Jiang, X., Hu, Y., Yao, J., Wang, Y., Xie, W.: Lamra: Large multimodal model as your advanced retrieval assistant. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4015–4025 (2025)
31. Meng, R., Jiang, Z., Liu, Y., Su, M., Yang, X., Fu, Y., Qin, C., Chen, Z., Xu, R., Xiong, C., et al.: Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. arXiv preprint arXiv:2507.04590 (2025)
32. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A human generated machine reading comprehension dataset. CoRR **abs/1611.09268** (2016), http://arxiv.org/abs/1611.09268
33. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
34. Radenović, F., Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Revisiting oxford and paris: Large-scale image retrieval benchmarking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5706–5715 (2018)
35. Sun, Q., Wang, J., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, X.: Eva-clip-18b: Scaling clip to 18 billion parameters. arXiv preprint arXiv:2402.04252 (2024)
36. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
37. Wang, H., Wang, Y., Zhang, T., Zhou, Y., Li, Y., Wang, J., Zheng, J., Tian, Y., Meng, J., Huang, Z., et al.: Grasp any region: Towards precise, contextual pixel understanding for multimodal llms. arXiv preprint arXiv:2510.18876 (2025)
38. Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., Wei, F.: Text embeddings by weakly-supervised contrastive pre-training. arXiv preprint arXiv:2212.03533 (2022)
39. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
40. Wei, C., Chen, Y., Chen, H., Hu, H., Zhang, G., Fu, J., Ritter, A., Chen, W.: Uniir: Training and benchmarking universal multimodal information retrievers. In: ECCV (2024)

41. Xiao, R., Kim, S., Georgescu, M.I., Akata, Z., Alaniz, S.: Flair: Vlm with fine-grained language-informed image representations. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24884–24894 (2025)
42. Xie, C., Wang, B., Kong, F., Li, J., Liang, D., Zhang, G., Leng, D., Yin, Y.: Fg-clip: Fine-grained visual and textual alignment. In: Proceedings of the International Conference on Machine Learning (2025)
43. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv preprint arXiv:2310.11441 (2023)
44. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W.W., Salakhutdinov, R., Manning, C.D.: HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2018)
45. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. arXiv preprint arXiv:2111.07783 (2021)
46. Yu, R., Ma, X., Wang, X.: Introducing visual perception token into multimodal large language model (2025)
47. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV (2023)
48. Zhang, B., Zhang, P., Dong, X., Zang, Y., Wang, J.: Long-clip: Unlocking the long-text capability of clip. In: European Conference on Computer Vision (2024)
49. Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Liu, Y., Chen, K., Luo, P.: Gpt4roi: Instruction tuning large language model on region-of-interest (2025), https://arxiv.org/abs/2307.03601
50. Zhang, X., Zhang, Y., Xie, W., Li, M., Dai, Z., Long, D., Xie, P., Zhang, M., Li, W., Zhang, M.: Gme: Improving universal multimodal retrieval by multimodal llms. arXiv preprint arXiv:2412.16855 (2024)
51. Zhang, Y., Pan, P., Zheng, Y., Zhao, K., Zhang, Y., Ren, X., Jin, R.: Visual search at alibaba. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 993–1001 (2018)
52. Zheng, K., Zhang, Y., Wu, W., Lu, F., Ma, S., Jin, X., Chen, W., Shen, Y.: Dreamlip: Language-image pre-training with long captions. In: European Conference on Computer Vision. pp. 73–90. Springer (2024)
53. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16793–16803 (2022)

Appendices

A Incorporating RoIs into Candidates

A key observation in region-level retrieval is the necessity of providing an explicit indicator, such as a RoI, to guide the model's attention to the relevant area. Without such guidance, the model often relies on global visual similarity, leading to suboptimal retrieval performance.

For instance, in Fig. 5(d), the negative candidates retrieved by LamRA share a dominant visual pattern: red vehicles with white stripes occupying most of the image. This suggests that the model is aligning the text with the full image rather than focusing on the specific region that better matches the query. This is because the model has to represent the full image with a single embedding when there are no extra modules to represent the RoI. The global-level embedding inevitably overlooks some local features and only utilizes the dominant visual cues as candidates. In the LamRA's semantic space, the true positive candidate image is closer to *"Two cars in front of a building"*, instead of *"A red van with white strips"*. Another example is in Fig. 5(e). The top-ranked negative candidate shares not only a similar outfit with the query but also a comparable background, including a horizon line located at one-third of the image height. However, this background similarity is misleading. The retrieval task only requires consistency in the foreground region, while the background, which should ideally have a café and a menu board, can differ from the query. Therefore, the standard approach to address these challenges is to include RoIs in both query and candidate inputs [5, 51]. This allows the model to focus on the relevant region. At the same time, background context can be downweighted, reducing the risk of interference from irrelevant visual features.

B Inference Time Analysis

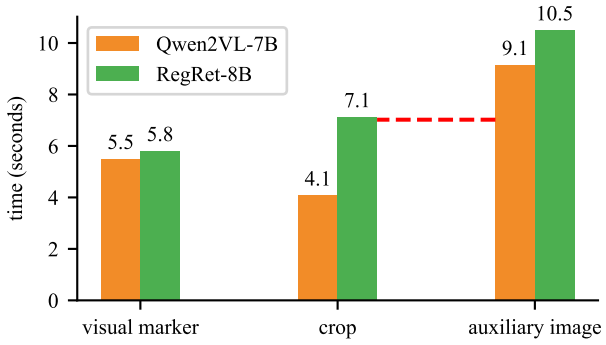

Fig. 6: Comparison of inference time with Qwen2VL-7B. The average time for generating a batch of embeddings on the I2I task of REGMB is reported. Compared with auxiliary image, which achieves the best average performance on REGMB, RegRet is 22% faster that baseline models.

To quantitatively assess the additional inference time introduced by incorporating RAE, we compare its speed with the base model, Qwen2VL-7B, under different prompting strategies. The

evaluation is conducted on samples with varying image sizes and target region scales. We report the average inference time per batch in Fig. 6, using a batch size of 96. Under the visual marker strategy, both RegRet and Qwen2VL-7B process the full image only once, resulting in comparable inference times. Under the crop strategy, RegRet is slower than Qwen. This slowdown occurs because RegRet encodes the entire image using the CE and also processes the RoI with RAE. In comparison, Qwen2VL-7B encodes only the RoI with CE, which is typically a small region, with considerably fewer visual tokens under Qwen's dynamic resolution paradigm. Meanwhile, given that RegRet achieves more than 20% performance improvement over Qwen-based baselines using the crop strategy, the additional inference cost is acceptable.

As to the auxiliary image strategy, this performance gap is narrowed. In this setting, both models encode two images: Qwen2VL-7B uses CE for both, while RegRet uses CE for the full image and RAE for the RoI. In practice, under the scenarios, such as task 4 on REGMB, where Qwen needs the auxiliary image strategy to fully boost their performance, RegRet is even faster (9.1⟶7.1), as illustrated in Fig. 6 with the red dashed line.

C Training Details

C.1 Training Parameters

In the first stage, we use the batch size 64 and the learning rate of 1×10^{-4}. Only RAE's cross-attention modules and the projector are trainable. In the second stage, the batch size is 1104 and the learning rate is 1.1×10^{-4}. We only update the language model backbone's parameters for one epoch using pure-text contrastive pairs. In the final stage, we fine-tune the language model backbone for two epochs on multimodal contrastive pairs, with a learning rate of 2×10^{-4} and a batch size of 1104. To save the GPU memory for a larger batch size, which is crucial in contrastive learning, we use LoRA to fine-tune the LLM. The LoRA rank is set to be 64 and 128 for stage two and stage three, respectively. All three stages can be completed within 20 hours on 24 NVIDIA H20 GPUs.

C.2 Training Strategy Ablations

Table 7: Ablations on the training strategy in Stage-III using 3B models, evaluated on REGMB.

Strategy	**Task1**	**Task2**	**Task3**	**Task4**	**Avg.**
crop	68.2	78.4	89.3	77.9	78.4
concat	68.3	77.3	88.5	71.4	75.8
random	67.8	78.9	89.2	73.0	76.7
mixed visual prompts	71.0	77.4	90.0	80.8	79.9

During the training stage three, we develop the *mixed visual prompting strategy* to balance contextual and regional information. In particular, for tasks 1 to 3 in the REGMB training split, which rely more on local details, we provide only RAE tokens to the LLM. For task 4, which requires contextual understanding, we concatenate CE tokens with RAE tokens. This design allows

the model to leverage CE tokens for context without injecting excessive background signals into RAE, thereby preserving fine-grained representation quality.

We compare the mixed visual prompting strategy with some approaches in 7. This strategy proves to be able to provide consistent performance improvements across both meta tasks. As shown in Table 7, we compare several alternative training strategies. The "crop" setting uses only the RAE tokens for both meta tasks; the "concat" setting uses the concatenation of context encoder (CE) and RAE tokens as visual input; and the "random" strategy randomly selects either "crop" or "concat" with a 50% probability for each meta task. We observe that our strategy achieves the best performance. This is mainly because of the trade-off between the two metatasks. Specifically, metatask 2 requires contextual information for comprehensive representation, while metatask 1 demands a highly localized, fine-grained understanding. Excessive context may thus act as a distraction in the latter. Therefore, we provide the CE tokens when training on metatask2, so that the model can access the global representations. Meanwhile, we use only RAE tokens on metatask 1 to prevent introducing background noise.

D Public Evaluation Benchmarks

In this section, we introduce the public datasets used in our experiments. As publicly available datasets with ROI annotations are relatively scarce, we carefully select three widely used, representative benchmarks with standard retrieval evaluation protocols for performance testing. Since most of these benchmarks are designed for image-to-image (I2I) retrieval, we apply a unified cropping strategy for all baseline models to ensure fair comparison, as discussed in Sec. 5.2.

$\mathcal{R}$Oxford. This dataset covers building scenes in Oxford, and is split into easy and hard subsets, with ROIs provided only for query images. In this work, we report results on the hard subset. We follow the original evaluation protocol and use mean Average Precision (mAP) as the metric. The top-50 metric fails to reflect actual retrieval performance here, because each query is associated with a large number of junk samples.

DeepFashion2. DeepFashion2 is a multi-task fashion dataset. We adopt its consumer-to-in-shop retrieval task, which aims to retrieve the most similar clothing items from the in-shop gallery based on user-uploaded consumer photos. Since the official test set candidate pool is not publicly available, we follow common practice and run experiments on the validation set. We report average Recall@1, which measures whether the positive candidate for each query outranks all negative samples. This is a stricter metric than mAP for this task.

ILIAS. This benchmark supports region-level retrieval on general images, covering both I2I and image-to-text (I2T) tasks. The original dataset includes a 1M distractor pool to increase retrieval difficulty. As indexing 1M candidates is excessively time-consuming, we use the 4.72k core subset for evaluation, and follow the original paper to adopt mAP@50 as the metric.

E Dataset Details

This section details the construction of REGMB. The benchmark is aggregated from a total of six data sources. These include four publicly available datasets-SAM, COYO, ImgDiff, and Vismin-and two proprietary datasets: XGoods and XLife, which we collected from a social media platform to cover a broader range of fused-modal scenarios. Using the bounding box annotations from these datasets, we created regional query-candidate pairs. To guarantee that the benchmark is sufficiently challenging and discriminative, each query is paired with one to six hard negatives across all tasks.

XGoods and XLife. Recognizing the limited availability of public datasets for region-level image-to-image retrieval, we collected XGoods from an social media platform. Unlike datasets such as NIGHTS, which only contain the same object in similar backgrounds, XGoods offers more complex and realistic scenes. In this task, a commercial product image acts as the query, while a user-uploaded photo of the same item serves as the positive candidate. We use a fine-tuned YOLO v8 model to detect the key objects, and use the regional feature to retrieve similar images using SigLIP2. Then we employ human annotators to select the relevant samples as positive candidates, and keep the rest of them as hard negatives. We paired each query with six highly relevant products as hard negatives. For Task 3, we allocated 69k samples for training and 3.3k for evaluation. For Task 4, we also built a 3.4k data test split as a complementary dataset.

Similar to XGoods, XLife was also collected from social media content on the same platform. The main difference is its focus on images from users' daily lives rather than commercial product displays. The data was processed using the same procedure as XGoods, from which we carefully selected 4.2k query-candidate pairs for the final test set. Please refer to Fig. 7 for the prompt used for generating descriptions.

SAM. This dataset is used to build Tasks 1 and 2. We first converted the segmentation masks from SAM into bounding boxes by taking the maximum coordinates along each axis, and discard those too small regions. Each region was generated using Qwen2.5-VL-72B with auxiliary image, as shown in Fig. 8. To ensure an adequate supply of hard negatives, we required that each candidate pool include at least two distinct regions. For both Task 1 and Task 2, this process yielded 30k samples for training and 1.5k for evaluation. We further manually checked the relevace of the true positive candidates in the evaluation dataset.

COYO. From the COYO dataset, we utilized bounding boxes previously generated by FG-CLIP and subsequently re-captioned each region with DAM. Employing much longer captions allows for a richer alignment between the image region and its associated semantic concepts. These region-caption pairs naturally form the basis for text-to-image and image-to-text retrieval tasks, for which we prepared 1.5k evaluation samples each.

ImgDiff. As an image editing dataset, each sample in ImgDiff contains a source image, an editing instruction, an edited image, and a bounding box indicating the modified area. This structure lends itself naturally to an image-text-to-image retrieval task, where the query is a composite of the source image and the instruction, and the positive candidate contains the target region. When constructing the test split, we exclusively retained samples with two or more editing instructions to serve as hard negatives. This split comprises 21k training samples and 1.3k test samples.

Vismin. The structure of Vismin is analogous to that of ImgDiff; the primary difference is that Vismin contains real-world photographs, whereas ImgDiff's images are in the AI-generated style. We applied the same processing approach: the source image and edit instruction from the query, and the modified image serves as the positive candidate. From this dataset, we selected 46k samples for training and 1.3k queries for testing.

Prompt

```
You are a contrastive learning data annotator skilled at describing
differences and relationships between images to generate query-candidate
pairs. Given an image of a query object and a set of candidate images,
for each candidate image, describe the environment where the query object
appears, its relationship with surrounding objects, or changes in its own
attributes. Examples:
a. A woman wearing this top is sitting in a studio broadcasting.
b. The same sofa is placed in a living room with a coffee table in front.
c. A green bottle of toner with a small cat sitting to its left.
d. A teddy bear of identical shape but brown instead of beige, placed among
a group of dolls.
Description rules:
a. Keep descriptions concise, under 100 words.
b. Include sufficient detail to clearly convey the specific context of the
query object in the candidate image.
c. Ensure uniqueness: the combination of the query and description should
uniquely identify the corresponding candidate image.
Output format:
a. Each candidate image receives one description and a score (between 0 and
1) indicating the relevance and accuracy of the description. Assign 0 if the
candidate is either too similar to the query or lacks distinctive contextual
details making description difficult; assign 1 if the description is clear
and highly discriminative. b. Descriptions must be in English.
c. Strictly follow JSON format as shown below:
"candidates": [
"0": "description0",
"score": "0.7"
,
"1": "description1",
"score": "0.1"
]
```

Fig. 7: The prompt for generating descriptions on Task 4, XGoods.

Prompt: Detailed Image Caption Generation

```
Describe the target region bounded by a red box in detail. Incorporate the
background to ensure a property understanding of the region.
```

Fig. 8: The prompt for generating regional captions.