MASK-GUIDED MULTI-GRANULARITY CONTRASTIVE LEARNING FOR FINE-GRAINED VISION–LANGUAGE ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive vision-language models such as CLIP excel at global image–text alignment but remain limited in fine-grained understanding. Existing methods construct local visual evidence by either implicitly selecting text-relevant patches or extracting region features with box-based operations such as RoIAlign. The former lacks explicit spatial grounding, while the latter often includes background or neighboring objects. To address these limitations, we introduce MaskAlign, a multi-granularity vision–language learning framework centered on pixel-level mask supervision. Masks provide precise spatial support, enabling the model to aggregate native visual patches predominantly covered by the target while suppressing background-dominated patches. We further construct MG-Data, comprising 6M images with long captions of over 100 words on average, short captions, and approximately 36M object and 14M region annotations, each pairing a curated mask with a detailed description. Using MG-Data, MaskAlign establishes image-, object-, and region-level visual–textual correspondences and integrates locally grounded semantics into global representations through cross-granularity alignment. We also introduce Long-caption Partial-Order (LPO), which promotes the use of additional details in long captions by enforcing a larger long-over-short similarity gain for matched pairs than for mined hard negatives. Extensive experiments demonstrate strong performance across long- and short-text retrieval, zero-shot classification, and fine-grained localization. An anonymous preview of our code is available at https://anonymous.4open.science/r/MaskAlign/.

1 INTRODUCTION

Models such as CLIP (Radford et al., 2021) learn shared visual–textual representations through global contrastive learning, achieving strong performance on image–text retrieval (Jia et al., 2021; Zhai et al., 2022; Li et al., 2023), zero-shot image classification (Conde & Turgutlu, 2021; Fan et al., 2023), and image captioning (Li et al., 2022a; Mokady et al., 2021).

However, these models (Radford et al., 2021; Xu et al., 2024; Zhai et al., 2023; Sun et al., 2023) remain limited on localized tasks such as bounding-box object recognition (Lin et al., 2014; Gupta et al., 2019) and semantic segmentation (Everingham et al., 2015; Cordts et al., 2016; Mottaghi et al., 2014; Zhou et al., 2019; Caesar et al., 2018). A global representation can act as an information bottleneck, obscuring the entities, attributes, and actions in detailed descriptions. Global alignment alone does not ensure that these semantics are locally grounded, integrated into global representations, or effectively used for discrimination. Addressing this challenge requires reliable local visual–textual correspondences that can enrich global representations.

Existing studies have mainly pursued two approaches to establishing such correspondences: unsupervised grounding and supervised localization. Unsupervised methods (Yao et al., 2022; Zheng et al., 2024; Bica et al., 2024; Asokan et al., 2025; Wu et al., 2024; Xiao et al., 2025; Zohra et al., 2026) discover text-relevant regions without direct spatial supervision, risking inaccurate correspondences. Supervised methods (Li et al., 2022b; Zhong et al., 2022; Xie et al., 2025; 2026) typically use bounding boxes, which support coarse localization but often include irrelevant background or neighboring instances.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address these limitations, we propose MaskAlign, a multi-granularity framework built on mask-guided local alignment. Each mask selects target-dominant native visual patches, whose features are averaged and aligned with the corresponding localized description. This explicit spatial support avoids implicit region discovery and reduces the background and neighboring content included within bounding boxes.

MaskAlign establishes intra-granularity visual–textual correspondences at three complementary granularities. Global image features are aligned with short captions capturing core semantics and long captions providing richer descriptions. Mask-guided object features are aligned with attribute-rich descriptions, while multi-object regions are aligned with descriptions of their shared context.

Although intra-granularity alignment directly grounds object- and region-level semantics, these local objectives do not directly constrain global image-text matching. We therefore introduce **cross-granularity alignment**, which aligns aggregated object- and region-level visual features with the global long caption, while aligning the corresponding aggregated textual features with the global image. This transfers locally grounded semantics into the global embedding space.

Complementary to cross-granularity alignment, **Long-caption Partial-Order (LPO)** promotes discrimination using additional long-caption details by enforcing a larger long-over-short similarity gain for matched pairs than for in-batch hard negatives mined via short-caption similarities.

To support this framework, we construct MG-Data with 6M images, short captions, long captions averaging over 100 words, and approximately 36M object and 14M region annotations, each pairing a curated mask with a localized description. Our main contributions are as follows:

- We propose MaskAlign, a mask-guided multi-granularity alignment framework. Masks provide precise spatial support for local visual representations, enabling image-, object-, and region-level visual–textual alignment. Cross-granularity alignment further integrates locally grounded semantics into global representations.
- We introduce Long-caption Partial-Order (LPO) to improve the discriminative use of long-caption details beyond the core semantics captured by short captions.
- We construct MG-Data, a large-scale dataset providing detailed image-level captions and approximately 50M object- and region-level mask–caption annotations for multi-granularity vision–language alignment.

2 Related Work

Contrastive Language–Image Pre-training CLIP (Radford et al., 2021) learns transferable visual–textual representations through large-scale, dual-encoder contrastive training. Its representations support image classification (Deng et al., 2009; Recht et al., 2019), image–text retrieval (Lin et al., 2014; Young et al., 2014; Urbanek et al., 2024; Chen et al., 2024), and open-vocabulary and zero-shot semantic segmentation (Lüddecke & Ecker, 2022; Dong et al., 2023; Liang et al., 2023; Lan et al., 2024; Wang et al., 2024; Xiao et al., 2025), motivating finer localized correspondences beyond global alignment.

Fine-Grained Exploration Long-CLIP (Zhang et al., 2024) extends global alignment to longer captions while preserving short-text knowledge. LoTLIP (Wu et al., 2024) introduces corner tokens to aggregate diverse textual information for long-text understanding. FineLIP (Asokan et al., 2025) dynamically aggregates local image and text tokens for fine-grained cross-modal alignment. Without explicit spatial supervision, FILIP (Yao et al., 2022) and DreamLIP (Zheng et al., 2024) derive local correspondences from visual–textual similarity, while FLAIR (Xiao et al., 2025) and β-CLIP (Zohra et al., 2026) use text-conditioned attention to pool relevant visual patches. In parallel, GLIP (Li et al., 2022b), RegionCLIP (Zhong et al., 2022), FG-CLIP (Xie et al., 2025), and FG-CLIP 2 (Xie et al., 2026) introduce localized supervision through bounding boxes. However, these boxes often include irrelevant background or neighboring objects, introducing noise into local alignment. For local feature extraction, the MLLM Ferret (You et al., 2024) offers a spatial-aware visual sampler, whose continuous region features are combined with discrete coordinates in a hybrid representation. These approaches obtain local features through text-conditioned attention pooling, box-based extraction, or spatial-aware sampling and aggregation within a region. MaskAlign selects target-dominant native visual patches based on mask coverage and averages their features. We com-

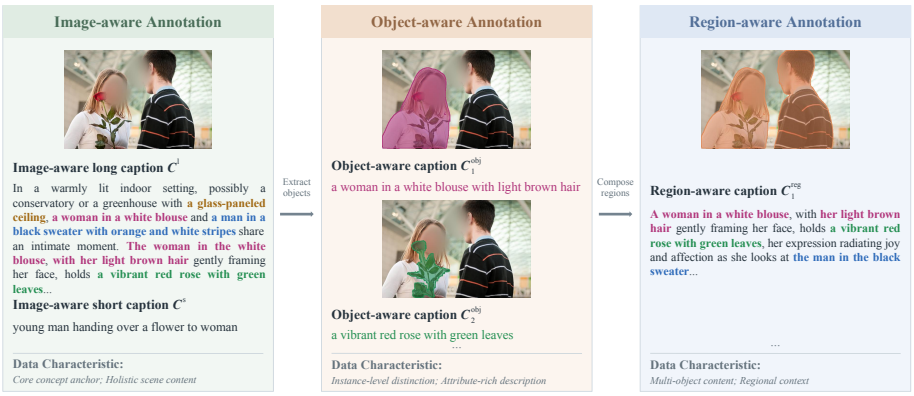

Figure 1: Multi-granularity annotations in MG-Data. Image-aware annotations include short and long captions; colors distinguish object and region mask instances.

pare these feature construction strategies under a unified training setting. Beyond local grounding, we further study how local semantics enrich global representations and how additional long-caption details improve discrimination.

3 Dataset Curation

Existing image–text datasets predominantly contain short captions; even long captions often describe the overall scene without elaborating on individual objects and their fine-grained details. Localized supervision remains scarce and typically relies on bounding boxes that include irrelevant background, weakening the correspondence between local features and textual semantics.

To overcome these limitations, we construct MG-Data, a large-scale dataset that provides complementary supervision at image, object, and region levels. Image-aware annotations describe holistic scenes with short and detailed captions; object-aware annotations pair individual entities and their fine-grained attributes with masks; region-aware annotations pair semantically related object groups with combined masks and descriptions of their shared context.

As illustrated in Figure 1, we represent an annotation at granularity $g \in \{img, obj, reg\}$ as $d = (I, M^g, C^g)$, where I is the image, M^g its spatial mask, and C^g the corresponding description. For image-aware annotations, M^{img} covers the full image and C^{img} comprises a short caption C^s and a long caption C^l. At the object and region levels, M^{obj} and M^{reg} specify the localized visual support for descriptions C^{obj} and C^{reg}, respectively. The following subsections outline the annotations at each granularity. Detailed annotation generation procedures are provided in Appendix A, with post-processing described in Appendix A.4.

3.1 Image-Aware Annotations

Image-aware annotations capture the holistic scene through complementary short and long captions. The short caption C^s serves as a concise "core concept anchor", while the long caption C^l adds visual attributes and contextual details for a more comprehensive description.

We adopt the source dataset's original captions as C^s. To generate C^l, we use a targeted LVLM prompt that elicits object attributes, spatial relationships, and contextual interactions, rather than generic requests such as "describe the image in detail". These detailed captions provide the basis for extracting object-aware and region-aware annotations.

3.2 Object-Aware Annotations

Whereas image-aware annotations describe the holistic scene, object-aware annotations provide explicit supervision for individual entities and their fine-grained attributes. We use an LLM to extract object captions C^{obj} from C^l, retaining attributes such as color, texture, and shape. These descrip-

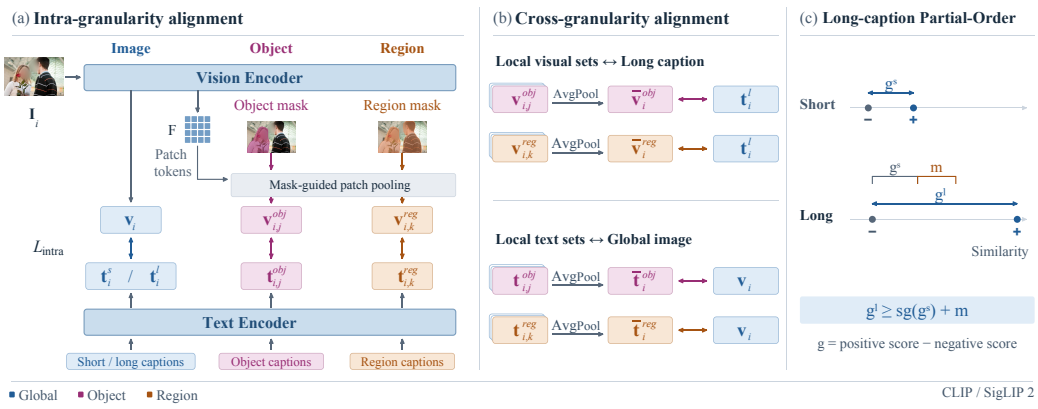

Figure 2: Overview of MaskAlign. **(a)** Intra-granularity alignment matches global image features and mask-pooled object/region features with their captions. **(b)** Cross-granularity alignment pairs aggregated local visual features with long captions and aggregated local textual features with global images. **(c)** LPO requires larger long-over-short similarity gains for matched pairs than for hard negatives, using short-caption similarities as stop-gradient references.

tions guide an open-vocabulary segmentation model to generate masks M^{obj}, grounding each entity to its spatial extent. Object masks further provide finer spatial support than bounding boxes by following target boundaries and excluding more background or neighboring objects. Table 5 evaluates how this spatial support affects downstream alignment.

3.3 Region-Aware Annotations

Beyond individual objects, detailed captions often describe multiple semantically related entities within a shared local context. We therefore use an LLM to parse each long caption, identify groups of relevant objects, and extract a region-aware caption C^{reg} describing their shared regional context. The corresponding region mask M^{reg} is constructed as the union of the masks of all involved objects.

These annotations complement object-level supervision by introducing multi-object regional units at an intermediate granularity between individual objects and the full image. Together, the object-, region-, and image-aware annotations provide complementary supervision spanning individual entities, shared local contexts, and holistic scenes.

4 Methodology

In this section, we present **MaskAlign**, a mask–guided multi-granularity vision–language alignment framework (Figure 2). Specifically, we first establish direct visual–textual correspondences at the image, object, and region levels through intra-granularity alignment. We then introduce cross-granularity alignment to leverage the grounded object- and region-level semantics to enrich global representations. Finally, we introduce Long-caption Partial-Order (LPO) to encourage additional fine-grained details to improve discrimination among semantically similar samples.

4.1 Intra-Granularity Alignment

Global image–text matching may rely on prominent scene concepts while overlooking individual entities, attributes, or multi-object content. We therefore use MG-Data to establish direct visual–textual correspondences at the image, object, and region levels.

At the image level, the vision encoder maps I_i to a global embedding $\mathbf{v}_i$, while the text encoder maps its short and long captions C_i^s and C_i^l to $\mathbf{t}_i^s$ and $\mathbf{t}_i^l$. We adopt the knowledge-preserving stretching strategy of Long-CLIP (Zhang et al., 2024) to accommodate long captions.

At the local level, MG-Data pairs object masks $M_{i,j}^{obj}$ with captions $C_{i,j}^{obj}$ describing individual entities and their attributes. Region masks $M_{i,k}^{\text{reg}}$ cover semantically related objects, with captions

$C^{\text{reg}}_{i,k}$ describing their shared context. The text encoder maps the object and region captions to $\mathbf{t}^{obj}_{i,j}$ and $\mathbf{t}^{reg}_{i,k}$, respectively. Their visual counterparts are extracted from the feature map $\mathbf{F}$ using patch pooling:

$$\mathbf{v}^{\text{obj}}_{i,j} = \text{PatchPooling}(\mathbf{F}, M^{\text{obj}}_{i,j}), \quad \mathbf{v}^{\text{reg}}_{i,k} = \text{PatchPooling}(\mathbf{F}, M^{\text{reg}}_{i,k}). \tag{1}$$

Patch pooling (Figure 3) selects visual features according to their mask coverage. For the image patch $P_{u,v}$ at grid location (u, v), coverage by mask M is computed as:

$$\alpha_{u,v} = \frac{1}{|P_{u,v}|} \sum_{(x,y)\in P_{u,v}} M(x, y), \tag{2}$$

where the mask is spatially aligned with the visual input. We retain patch locations whose mask coverage meets the threshold $\alpha_{\text{adp}} = 0.5$, forming $\mathcal{S}(M) = \{(u, v) \mid \alpha_{u,v} \geq \alpha_{\text{adp}}\}$, and uniformly average their features:

$$\text{PatchPooling}(\mathbf{F}, M) = \frac{1}{|\mathcal{S}(M)|} \sum_{(u,v)\in\mathcal{S}(M)} \mathbf{F}_{u,v}. \tag{3}$$

The precise mask shape suppresses patches dominated by background or neighboring instances, while direct feature gathering preserves the native visual tokens without spatial interpolation. Moreover, the selected support is explicitly determined by the mask rather than implicitly inferred through a text-conditioned pooling module, providing spatially grounded local evidence for subsequent alignment.

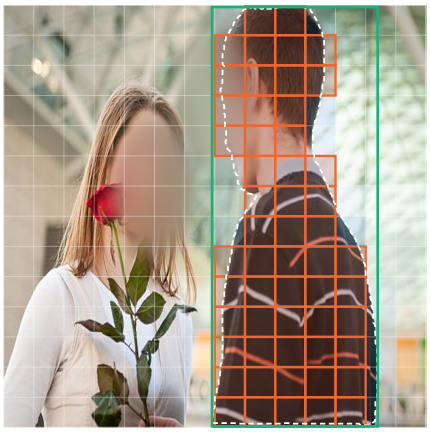

Figure 3: Mask-guided patch pooling. The white contour marks the mask, the green rectangle its enclosing box, and the grid the visual patches. Patches with $\alpha_{u,v} \geq \alpha_{\text{adp}}$ are selected and averaged.

For a batch of paired visual and textual representations $\{(\mathbf{a}_i, \mathbf{b}_i)\}_{i=1}^{N}$ at a given granularity, we use the bidirectional contrastive loss:

$$\mathcal{L}_{\text{align}}(\mathbf{a}, \mathbf{b}) = -\frac{1}{2N} \sum_{i=1}^{N} \left[\log \frac{\exp(\text{sim}(\mathbf{a}_i, \mathbf{b}_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{a}_i, \mathbf{b}_j)/\tau)} + \log \frac{\exp(\text{sim}(\mathbf{b}_i, \mathbf{a}_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{b}_i, \mathbf{a}_j)/\tau)} \right], \tag{4}$$

where $\text{sim}(\cdot, \cdot)$ is cosine similarity and τ is the learnable temperature. For the short- and long-caption image pairs, the image-level objective is

$$\mathcal{L}_{\text{img}} = \mathcal{L}_{\text{align}}(\mathbf{v}, \mathbf{t}^s) + \mathcal{L}_{\text{align}}(\mathbf{v}, \mathbf{t}^l). \tag{5}$$

Applying the same bidirectional contrastive loss to valid pairs within the object and region granularities gives $\mathcal{L}_{\text{obj}}$ and $\mathcal{L}_{\text{reg}}$, respectively. The combined intra-granularity objective is:

$$\mathcal{L}_{\text{intra}} = \mathcal{L}_{\text{img}} + \lambda_o \mathcal{L}_{\text{obj}} + \lambda_r \mathcal{L}_{\text{reg}}, \tag{6}$$

where $\lambda_o = \lambda_r = 0.1$ in all experiments.

4.2 Cross-Granularity Alignment

Intra-granularity alignment grounds descriptions at their corresponding visual scales, but it does not by itself ensure that the localized semantics are reflected in the global representations used for image–text matching. An image may, for example, contain objects and regional context that are correctly aligned with their local captions, while its global embedding still emphasizes only the most prominent scene concepts. We therefore connect local and global representations through cross-granularity alignment.

For each image, we aggregate its valid object- and region-level visual features into $\bar{\mathbf{v}}^{\text{obj}}_i$ and $\bar{\mathbf{v}}^{\text{reg}}_i$, respectively. We likewise aggregate the corresponding textual features into $\bar{\mathbf{t}}^{\text{obj}}_i$ and $\bar{\mathbf{t}}^{\text{reg}}_i$:

$$\bar{\mathbf{v}}^{obj}_i = \frac{1}{N_o} \sum_{j=1}^{N_o} \mathbf{v}^{obj}_{i,j}, \quad \bar{\mathbf{t}}^{obj}_i = \frac{1}{N_o} \sum_{j=1}^{N_o} \mathbf{t}^{obj}_{i,j}, \quad \bar{\mathbf{v}}^{reg}_i = \frac{1}{N_r} \sum_{k=1}^{N_r} \mathbf{v}^{reg}_{i,k}, \quad \bar{\mathbf{t}}^{reg}_i = \frac{1}{N_r} \sum_{k=1}^{N_r} \mathbf{t}^{reg}_{i,k}, \tag{7}$$

where N_o and N_r denote the numbers of valid objects and regions in image i, respectively.

The alignment operates in two complementary directions. First, we align the aggregated object- and region-level visual features with the long-caption representation $\mathbf{t}^l_i$. These alignments encourage

the global textual representation to capture locally grounded entity and attribute semantics, as well as multi-object regional context. Second, we align the global image representation $\mathbf{v}_i$ with the aggregated object- and region-level textual features, encouraging it to incorporate the fine-grained semantics expressed in the localized descriptions. The cross-granularity pairs are optimized with the same bidirectional contrastive objective:

$$\mathcal{L}_{\text{cross}} = \mathcal{L}_{\text{align}}(\bar{\mathbf{v}}^{obj}, \mathbf{t}^{l}) + \mathcal{L}_{\text{align}}(\bar{\mathbf{v}}^{reg}, \mathbf{t}^{l}) + \mathcal{L}_{\text{align}}(\mathbf{v}, \bar{\mathbf{t}}^{obj}) + \mathcal{L}_{\text{align}}(\mathbf{v}, \bar{\mathbf{t}}^{reg}). \tag{8}$$

Cross-granularity alignment thus complements intra-granularity alignment by connecting locally grounded semantics with the global representations used for image–text matching.

4.3 Detail-Sensitive Global Discrimination

Cross-granularity alignment enriches global representations with locally grounded semantics. Complementing this alignment, we introduce Long-caption Partial-Order (LPO) to encourage the additional entities, attributes, and contextual details in long captions to improve discrimination beyond the core concepts captured by short captions. LPO mines in-batch hard negatives using short-caption similarities and requires the long-over-short similarity gain to be larger for matched pairs than for the mined negatives.

Let $s(I, C)$ denote the logit-scaled cosine similarity between global image and caption representations, and $\text{sg}(\cdot)$ the stop-gradient operator. For image-to-text retrieval, we measure the long-over-short similarity gain, treating the short-caption score as a stop-gradient reference for LPO:

$$\Delta_{i,j}^{I2T} = s(I_i, C_j^{l}) - \text{sg}\big(s(I_i, C_j^{s})\big). \tag{9}$$

For each image I_i, we select the top-K mismatched captions with the highest short-caption similarities to I_i, forming the set $\mathcal{H}_i^{I2T}$. We require the similarity gain for the matched caption to exceed that for each hard-negative caption by a margin $m = 0.05$:

$$\mathcal{L}_{\text{lpo}}^{I2T} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{j \in \mathcal{H}_i^{I2T}} \left[m - \left(\Delta_{i,i}^{I2T} - \Delta_{i,j}^{I2T} \right) \right]_+, \tag{10}$$

where $[x]_+ = \max(x, 0)$. The text-to-image loss $\mathcal{L}_{\text{lpo}}^{T2I}$ is defined symmetrically, using the top-K mismatched images with the highest similarities to each short-caption query C_i^s and applying the same margin constraint. The final LPO objective is:

$$\mathcal{L}_{\text{lpo}} = \mathcal{L}_{\text{lpo}}^{I2T} + \mathcal{L}_{\text{lpo}}^{T2I}. \tag{11}$$

Unlike a conventional ranking loss on long-caption similarities, LPO requires the long-over-short similarity gain to be at least m larger for matched pairs than for hard negatives. Let g_l and g_s denote the long- and short-caption positive–negative similarity gaps. Each term is equivalently $[m - g_l + \text{sg}(g_s)]_+$. The stop-gradient operation prevents LPO from directly weakening short-caption discrimination to satisfy the constraint.

4.4 Overall Objective

The three objectives are combined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{intra}} + \lambda_c \mathcal{L}_{\text{cross}} + \lambda_p \mathcal{L}_{\text{lpo}}, \tag{12}$$

where the loss weights are fixed to $\lambda_c = 0.1$ and $\lambda_p = 0.5$ throughout training.

5 Experiments

5.1 Experimental Setting

Models We evaluate CLIP (Radford et al., 2021), EVA-CLIP (Sun et al., 2023), Long-CLIP (Zhang et al., 2024), FineCLIP (Jing et al., 2024), MetaCLIP (Xu et al., 2024), SigLIP 2 (Tschannen et al., 2025), β-CLIP (Zohra et al., 2026), FG-CLIP (Xie et al., 2025), and FG-CLIP 2 (Xie et al., 2026). For controlled comparisons, we retrain β-CLIP, FG-CLIP, and FG-CLIP 2 on MG-Data following their original training setups, converting masks to minimum enclosing boxes for the two FG-CLIP models. To keep supervision restricted to MG-Data, we do not introduce method-specific hard-negative samples during retraining, consistent with the FG-CLIP ablation showing that such supervision reduces performance on long- and short-text retrieval and bounding-box classification.

Table 1: Bounding-box/mask classification Top-1 accuracy, reported as bbox/mask. Best and second-best results are **bold** and underlined.

Method	ViT-B		ViT-L	
	COCO	LVIS	COCO	LVIS
CLIP	44.2/45.5	20.9/21.7	33.8/34.9	9.3/10.0
EVA-CLIP	30.6/31.2	14.4/14.9	32.1/32.9	18.3/18.7
Long-CLIP	36.7/37.4	18.2/18.9	35.6/36.6	10.4/11.2
FineCLIP	48.4/49.3	23.3/24.0	54.5/55.2	22.5/23.3
SigLIP 2	28.0/28.0	17.5/17.4	28.1/28.1	18.1/18.0
β-CLIP	**59.0/58.6**	18.8/19.8	37.2/37.5	5.4/5.8
FG-CLIP	50.2/49.8	**24.8**/24.2	60.9/60.9	**33.9/33.2**
FG-CLIP 2	28.1/28.0	13.1/13.1	30.3/30.6	20.3/20.4
MaskAlign	57.6/58.0	22.7/23.2	**62.5/63.8**	33.0/33.0
MaskAlign (SigLIP 2)	39.4/40.7	23.9/**24.6**	41.4/42.5	26.1/26.8

Table 2: Zero-shot semantic segmentation (mIoU) on VOC20 (VOC), Cityscapes (City), Context59 (Con), ADE20K (ADE), and COCO-Stuff (COCO), with ViT-B/16 for all models.

Method	VOC	City	Con	ADE	COCO
CLIP	41.8	5.5	9.2	3.2	4.4
MetaCLIP	35.4	5.0	8.1	2.2	4.3
SigLIP 2	73.7	12.7	13.9	7.0	8.4
Long-CLIP	37.4	5.7	8.8	2.0	4.6
β-CLIP	56.2	13.2	21.8	10.1	13.5
FG-CLIP	59.2	12.0	23.0	10.6	14.3
FG-CLIP 2	74.3	14.7	17.4	7.6	10.4
MaskAlign	66.0	13.7	25.2	12.3	16.1
MaskAlign (SigLIP 2)	**78.1**	**20.2**	**30.0**	**15.5**	**19.8**

Implementation Details We train CLIP and SigLIP 2 with ViT-B and ViT-L backbones for one epoch on MG-Data, initializing each from its corresponding pretrained weights. We use AdamW with a learning rate of 3×10^{-6} and a weight decay of 1×10^{-3}. The global batch size is 4,096, except for SigLIP 2 ViT-L (3,072). We use GradCache (Gao et al., 2021) with chunk sizes of 256 and 128 for CLIP and SigLIP 2, respectively. Training employs DeepSpeed ZeRO-2 (Rajbhandari et al., 2020), TF32, gradient checkpointing, and BFloat16 mixed precision.

5.2 Region-based Image Understanding Tasks

Bounding Box and Mask Classification We begin our experimental evaluation by assessing the model's localized comprehension, examining whether the localized visual features align with their corresponding category text embeddings. Specifically, we apply RoIAlign (He et al., 2017) to extract bounding-box features and patch pooling to extract mask features. Following Jing et al. (2024), we evaluate the extracted features on the bbox and mask classification tasks and report Top-1 accuracy on COCO val2017 (Lin et al., 2014) and LVIS (Gupta et al., 2019).

As shown in Table 1, MaskAlign demonstrates strong localized classification performance across different backbones and model scales. With CLIP, MaskAlign outperforms FG-CLIP on COCO by +7.4/+8.2 percentage points with ViT-B and +1.6/+2.9 points with ViT-L on bounding-box/mask classification. With SigLIP 2, MaskAlign consistently outperforms FG-CLIP 2 on both datasets, with gains of at least +11.1/+11.9 points on COCO and +5.8/+6.4 points on LVIS across the two model scales.

Zero-Shot Semantic Segmentation To further evaluate the model's localized comprehension capabilities, we follow the evaluation protocols described in prior work (Lan et al., 2024; Wang et al., 2024; Xiao et al., 2025) to examine the alignment of visual patch features with their corresponding semantic labels. The evaluation is conducted on VOC20 (Everingham et al., 2015), Cityscapes (Cordts et al., 2016), Context59 (Mottaghi et al., 2014), ADE20K (Zhou et al., 2019), and COCO-Stuff (Caesar et al., 2018), reporting mIoU.

As shown in Table 2, MaskAlign outperforms FG-CLIP across all five datasets, with gains ranging from +1.7 to +6.8 mIoU. MaskAlign (SigLIP 2) further achieves the best performance among the compared methods on every benchmark, surpassing FG-CLIP 2 by +3.8 to +12.6 mIoU. These results demonstrate strong and consistent localized performance across both backbone families.

5.3 Holistic Image Understanding Tasks

Long/Short Caption Image-Text Retrieval We evaluate long-text retrieval on DCI (Urbanek et al., 2024), Urban1K (Zhang et al., 2024), and the 1K subset of ShareGPT4V (Chen et al., 2024), and short-text retrieval on the validation set of MSCOCO 5K (Lin et al., 2014) and the test set of Flickr30K (Young et al., 2014). As shown in Table 3, both MaskAlign variants consistently outperform their corresponding FG-CLIP counterparts on DCI, Urban1K, and ShareGPT4V across different model scales, while maintaining competitive performance on short-text retrieval.

Table 3: Comparisons on image-level tasks, including long/short caption image-text retrieval and zero-shot image classification. Recall@1 for retrieval and Top-1 accuracy for classification are reported, and the best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Backbone	DCI		Urban1K		ShareGPT4V		MSCOCO		Flickr30K		ImageNet-1K	ImageNet-v2
		I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	Top-1	Top-1
CLIP	ViT-B/16	45.5	43.0	68.1	53.6	78.2	79.6	51.8	32.7	82.2	62.1	68.4	61.9
EVA-CLIP	ViT-B/16	41.9	41.2	67.1	60.9	90.5	85.5	58.7	41.6	85.7	71.2	74.7	67.0
SigLIP 2	ViT-B/16	32.3	34.2	36.6	34.9	66.0	67.9	71.2	**55.2**	92.6	78.0	**81.2**	**74.5**
Long-CLIP	ViT-B/16	51.7	57.3	78.9	79.5	**94.7**	93.4	57.6	40.4	85.9	70.7	66.8	61.2
FineCLIP	ViT-B/16	35.5	34.4	62.5	65.2	70.6	73.3	54.5	40.2	82.5	67.9	55.7	48.8
β-CLIP	ViT-B/16	53.1	57.5	89.9	90.0	92.8	91.4	49.2	34.4	81.2	68.0	58.3	52.0
FG-CLIP	ViT-B/16	52.5	51.2	88.0	84.9	93.8	93.7	60.7	41.5	88.4	72.4	69.3	62.2
FG-CLIP 2	ViT-B/16	52.8	54.1	82.2	79.6	92.6	90.2	**72.5**	54.4	94.8	**83.3**	78.6	72.3
MaskAlign	ViT-B/16	58.4	**59.1**	**93.0**	**92.1**	**94.7**	**93.8**	60.9	41.8	89.4	73.6	68.1	61.8
MaskAlign (SigLIP 2)	ViT-B/16	**58.6**	58.9	88.2	87.3	94.6	92.3	72.1	54.8	**95.4**	**83.3**	79.6	72.8
CLIP	ViT-L/14	37.2	36.4	68.7	52.8	86.5	83.6	58.0	37.1	87.4	67.3	76.6	70.9
EVA-CLIP	ViT-L/14	47.2	47.8	76.6	70.1	91.5	89.4	64.2	47.9	89.2	77.9	80.4	73.8
SigLIP 2	ViT-L/16	46.7	47.6	63.1	60.8	84.9	82.8	71.5	54.7	94.5	**84.1**	**82.5**	**76.8**
Long-CLIP	ViT-L/14	44.2	52.5	82.7	86.1	95.8	**95.6**	62.8	46.3	90.0	76.2	73.5	67.9
FineCLIP	ViT-L/14	40.1	46.2	66.6	74.0	73.4	82.7	-	-	-	-	60.8	53.4
β-CLIP	ViT-L/14	58.5	65.0	**95.4**	**94.6**	95.3	94.2	52.4	39.7	88.0	75.8	66.3	60.4
FG-CLIP	ViT-L/14	58.3	59.7	93.2	91.4	95.6	92.7	65.6	46.4	92.4	78.9	76.8	71.1
FG-CLIP 2	ViT-L/16	60.6	61.7	88.2	85.5	94.5	92.4	72.4	55.9	**95.6**	83.4	82.2	75.8
MaskAlign	ViT-L/14	**65.3**	**67.3**	94.5	**94.6**	**96.1**	95.1	65.4	46.6	92.9	79.1	75.6	69.7
MaskAlign (SigLIP 2)	ViT-L/16	61.2	65.5	93.3	92.4	95.9	94.8	**73.1**	**56.2**	95.3	83.4	82.3	76.0

Table 4: Ablation of MaskAlign on CLIP and SigLIP 2 by progressively adding image-, object-, and region-level alignment, followed by $\mathcal{L}_{\text{cross}}$ and $\mathcal{L}_{\text{lpo}}$. Long and short retrieval report I2T/T2I Recall@1 on DCI and Flickr30K, respectively. Classification reports Top-1 accuracy on ImageNet-1K and Top-1/Top-5 accuracy for COCO bounding-box classification; VOC segmentation reports mIoU. Setting #4 corresponds to CLIP/SigLIP 2 + $\mathcal{L}_{\text{intra}}$.

Setting	Method	Long Retrieval		Short Retrieval		ImageNet-1K	BBox Cls.		VOC Seg.
		I2T	T2I	I2T	T2I		Top-1	Top-5	
#1	CLIP	45.5	43.0	82.2	62.1	68.4	44.2	72.3	41.8
#2	CLIP + $\mathcal{L}_{\text{img}}$	52.5	52.0	87.8	**73.6**	**69.5**	57.9	83.0	56.2
#3	CLIP + $\mathcal{L}_{\text{img}}$ + $\mathcal{L}_{\text{obj}}$	52.7	52.1	88.0	72.8	69.0	**58.9**	**85.7**	65.6
#4	CLIP + $\mathcal{L}_{\text{img}}$ + $\mathcal{L}_{\text{obj}}$ + $\mathcal{L}_{\text{reg}}$	53.5	53.0	88.4	72.6	68.9	57.1	85.0	**66.8**
#5	CLIP + $\mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{cross}}$	57.1	58.0	88.9	73.1	68.0	57.7	85.4	66.0
#6	CLIP + $\mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{cross}} + \mathcal{L}_{\text{lpo}}$	**58.4**	**59.1**	**89.4**	**73.6**	68.1	57.6	85.3	66.0
#1	SigLIP 2	32.3	34.2	92.6	78.0	**81.2**	28.0	66.8	73.7
#2	SigLIP 2 + $\mathcal{L}_{\text{img}}$	55.0	55.6	94.7	82.6	79.6	28.0	74.1	72.5
#3	SigLIP 2 + $\mathcal{L}_{\text{img}}$ + $\mathcal{L}_{\text{obj}}$	55.4	56.1	94.8	82.9	79.6	**40.1**	81.9	75.6
#4	SigLIP 2 + $\mathcal{L}_{\text{img}}$ + $\mathcal{L}_{\text{obj}}$ + $\mathcal{L}_{\text{reg}}$	56.5	57.0	95.2	83.0	79.7	39.1	80.6	77.6
#5	SigLIP 2 + $\mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{cross}}$	57.9	57.8	**95.4**	83.2	79.6	39.4	82.1	77.7
#6	SigLIP 2 + $\mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{cross}} + \mathcal{L}_{\text{lpo}}$	**58.6**	**58.9**	**95.4**	**83.3**	79.6	39.4	**82.4**	**78.1**

Zero-shot Image Classification As shown in Table 3, MaskAlign maintains competitive performance on ImageNet-1K (Deng et al., 2009) and ImageNet-v2 (Recht et al., 2019) across different backbones and model scales, indicating that fine-grained alignment does not substantially compromise the model's global recognition capability.

5.4 Ablation Studies

Effects of Different Components in MaskAlign Starting from the pretrained models, we progressively introduce image-, object-, and region-level alignment, followed by cross-granularity alignment $\mathcal{L}_{\text{cross}}$ and detail-sensitive global discrimination $\mathcal{L}_{\text{lpo}}$. As shown in Table 4, image-only training improves long- and short-text retrieval over the corresponding pretrained models. Adding object-level alignment improves VOC segmentation by +9.4/+3.1 mIoU for CLIP/SigLIP 2; region-level alignment provides a further +1.2/+2.0 mIoU gain and improves DCI retrieval on both backbones. Cross-granularity alignment further improves DCI retrieval by +3.6/+5.0 with CLIP and +1.4/+0.8 with SigLIP 2, demonstrating that local semantics can effectively benefit global representations. Finally, $\mathcal{L}_{\text{lpo}}$ brings additional gains on long-text retrieval, while maintaining comparable short-text retrieval, classification, and localized recognition performance. Appendix B.4 provides a qualita-

Table 5: Controlled comparison of local visual feature construction strategies. MSCOCO and DCI retrieval are reported as I2T/T2I Recall@1; bounding-box classification is reported as Top-1/Top-5 accuracy. IN-v2 reports Top-1 accuracy, and segmentation reports mIoU.

Backbone	Method	Image-Level Tasks			BBox Classification		Segmentation				
		MSCOCO	DCI	IN-v2	COCO	LVIS	VOC	Con	COCO	City	ADE
CLIP	RoIAlign	60.5/**42.1**	58.0/58.7	61.7	50.4/80.7	**25.0/47.7**	58.9	22.7	14.1	11.7	11.0
	RLE-BBox	60.7/42.0	58.2/59.0	61.6	51.1/81.6	24.6/47.2	60.6	23.1	14.5	12.0	11.3
	Ferret-style	60.8/41.2	57.8/57.7	61.0	**59.4**/84.0	22.7/43.8	60.5	24.4	15.2	13.3	10.7
	β-CLIP-style	60.4/41.6	57.1/57.3	61.6	58.3/83.1	20.6/40.3	50.9	20.5	13.3	11.2	9.3
	Mask (ours)	**60.9**/41.8	**58.4/59.1**	**61.8**	57.6/**85.3**	22.7/45.0	**66.0**	**25.2**	**16.1**	**13.7**	**12.3**
SigLIP 2	RoIAlign	71.5/**54.9**	58.3/58.9	**72.8**	36.3/70.6	22.4/46.7	73.3	24.5	16.0	14.9	12.1
	RLE-BBox	71.4/54.8	58.2/**59.0**	**72.8**	36.2/71.5	22.5/46.7	73.6	24.6	16.1	14.9	12.2
	Ferret-style	71.7/54.6	**58.7**/58.6	72.7	**41.2/84.0**	13.9/32.9	**80.8**	24.2	15.0	19.1	10.6
	β-CLIP-style	70.4/54.5	58.1/58.2	72.4	28.1/75.4	15.8/35.9	72.2	13.9	8.7	8.6	5.9
	Mask (ours)	**72.1**/54.8	58.6/58.9	**72.8**	39.4/82.4	**23.9/49.1**	78.1	**30.0**	**19.8**	**20.2**	**15.5**

tive analysis of $\mathcal{L}_{\text{lpo}}$. These results demonstrate the complementary effects of the three components across both model families.

Effect of Spatial Support for Local Visual Evidence We compare five strategies under identical data, alignment objectives, and training settings. **Mask** averages native patch features selected by pixel-level mask coverage. **RoIAlign** interpolates bounding-box features; **RLE-BBox** applies the same patch pooling as Mask to binary masks of minimum enclosing boxes. **β-CLIP-style** uses the official text-conditioned attention pooling module without spatial annotations. **Ferret-style** uses the complete hybrid region representation, combining discrete coordinates with continuous features from Ferret's spatial-aware visual sampler.

Mask vs. Box-Based Support Since Mask and RLE-BBox use the same patch pooling and differ only in spatial support, their comparison directly isolates the effect of precise mask geometry. Mask consistently improves all five segmentation benchmarks across both CLIP and SigLIP 2 while maintaining comparable image-level performance. It also consistently outperforms RoIAlign on segmentation, further demonstrating the benefit of precise mask support over box-based feature extraction. The training-mask perturbation study in Appendix B.3 further supports the importance of spatially accurate supervision.

Mask vs. Implicit Attention Compared with β-CLIP-style, Mask substantially improves segmentation across both backbones while maintaining strong retrieval and localized classification performance. These results support the benefit of explicit mask guidance for learning spatially precise local representations.

Mask vs. Ferret-Style Region Extraction Mask remains competitive with the complete Ferret-style hybrid extractor on retrieval and localized classification, while performing better on nearly all segmentation benchmarks. Ferret-style requires approximately 2-3$\times$ the training time of Mask under the same setting, highlighting the effectiveness and efficiency of mask-supported patch aggregation for spatially precise local representations.

6 Conclusion

In this paper, we constructed MG-Data, comprising image-aware captions and object- and region-aware mask-caption annotations. Building on this dataset, we introduced MaskAlign, a multi-granularity alignment framework that establishes visual–textual correspondences at the image, object, and region levels, with masks providing precise spatial support for localized alignment. We further introduced cross-granularity alignment to integrate locally learned fine-grained semantics into global representations, and Long-caption Partial-Order (LPO) to encourage the additional information in long captions to contribute to fine-grained discrimination. Extensive experiments demonstrate the effectiveness of MaskAlign across multiple downstream tasks and backbone architectures.

ETHICS STATEMENT

MG-Data is constructed from the publicly available CC3M and CC12M datasets, subject to their applicable access and usage terms. We acknowledge potential privacy, bias, and inappropriate-content concerns associated with web-sourced data and AI-generated annotations.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we describe the dataset construction pipeline, training objectives, and experimental settings in Sections 3–5 and Appendix A, with additional experiments in Appendix B. A preview of our code is available in an anonymous repository for review. The source datasets, CC3M and CC12M, are publicly available subject to their respective access and usage terms.

AI USE STATEMENT

AI models were used to generate annotations for MG-Data, followed by the filtering procedures described in Appendix A. AI tools also assisted with manuscript writing and polishing to improve clarity and readability. The authors reviewed and revised the AI-assisted manuscript text and take full responsibility for the final manuscript and reported results.

REFERENCES

Xiang An, Kaicheng Yang, Xiangzi Dai, Ziyong Feng, and Jiankang Deng. Multi-label cluster discrimination for visual representation learning. In *Proc. Eur. Conf. Comp. Vis.*, pp. 428–444, 2024.

Mothilal Asokan, Kebin Wu, and Fatima Albreiki. FineLIP: Extending CLIP's reach via fine-grained alignment with longer text inputs. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 14495–14504, 2025.

Ioana Bica, Anastasija Ilić, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A. Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pașcanu, and Jovana Mitrović. Improving fine-grained understanding in image-text pre-training. In *Proc. Int. Conf. Mach. Learn.*, pp. 3974–3995, 2024.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 1209–1218, 2018.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 3558–3568, 2021.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4V: Improving large multi-modal models with better captions. In *Proc. Eur. Conf. Comp. Vis.*, pp. 370–387, 2024.

Marcos V Conde and Kerem Turgutlu. CLIP-Art: Contrastive pre-training for fine-grained art classification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn. Workshops*, pp. 3956–3960, 2021.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 3213–3223, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 248–255, 2009.

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. MaskCLIP: Masked self-distillation advances contrastive language-image pretraining. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 10995–11005, 2023.

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes challenge: A retrospective. *Int. J. Comput. Vision*, 111(1):98–136, 2015.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving CLIP training with language rewrites. In *Advances in Neural Inf. Process. Syst.*, pp. 35544–35575, 2023.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. Scaling deep contrastive learning batch size under memory limited setup. In *Proc. Workshop on Representation Learning for NLP*, pp. 316–321, 2021.

Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 5356–5364, 2019.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proc. IEEE Int. Conf. Comp. Vis.*, pp. 2961–2969, 2017.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. Int. Conf. Mach. Learn.*, pp. 4904–4916, 2021.

Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Guoxing Yang, Wei Wei, Huiwen Zhao, and Zhiwu Lu. FineCLIP: Self-distilled region-based CLIP for better fine-grained understanding. In *Advances in Neural Inf. Process. Syst.*, pp. 27896–27918, 2024.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ClearCLIP: Decomposing CLIP representations for dense vision-language inference. In *Proc. Eur. Conf. Comp. Vis.*, pp. 143–160, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proc. Int. Conf. Mach. Learn.*, pp. 12888–12900, 2022a.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 10965–10975, 2022b.

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 23390–23400, 2023.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted CLIP. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 7061–7070, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. Eur. Conf. Comp. Vis.*, pp. 740–755, 2014.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 7086–7096, 2022.

Ron Mokady, Amir Hertz, and Amit H Bermano. CLIPCap: CLIP prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 891–898, 2014.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. Mach. Learn.*, pp. 8748–8763, 2021.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *SC20: International conference for high performance computing, networking, storage and analysis*, pp. 1–16. IEEE, 2020.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proc. Int. Conf. Mach. Learn.*, pp. 5389–5400, 2019.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proc. Conf. Association for Computational Linguistics*, pp. 2556–2565, 2018.

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

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating CLIP-style models on dense captions. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 26700–26709, 2024.

Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking self-attention for dense vision-language inference. In *Proc. Eur. Conf. Comp. Vis.*, pp. 315–332, 2024.

Kun Wu, Yin Xie, Yu Jie, Zhang Wenkang, Xinyu Zhou, Xiang An, Ziyong Feng, and Jiankang Deng. MLCD-Seg, 2025. URL `https://github.com/deepglint/MLCD-Seg`.

Wei Wu, Kecheng Zheng, Shuailei Ma, Fan Lu, Yuxin Guo, Yifei Zhang, Wei Chen, Qingpei Guo, Yujun Shen, and Zheng-Jun Zha. LoTLIP: Improving language-image pre-training for long text understanding. In *Advances in Neural Inf. Process. Syst.*, pp. 64996–65019, 2024.

Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. FLAIR: VLM with fine-grained language-informed image representations. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 24884–24894, 2025.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. FG-CLIP: Fine-grained visual and textual alignment. In *Proc. Int. Conf. Mach. Learn.*, pp. 68777–68793, 2025.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Ji Ao, Dawei Leng, and Yuhui Yin. FG-CLIP 2: A bilingual fine-grained vision-language alignment model. In *Proc. Int. Conf. Mach. Learn.*, 2026.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *Proc. Int. Conf. Learn. Representations*, pp. 47812–47831, 2024.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *Proc. Int. Conf. Learn. Representations*, 2022.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *Proc. Int. Conf. Learn. Representations*, pp. 57153–57180, 2024.

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

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-shot transfer with locked-image text tuning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 18123–18133, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proc. IEEE Int. Conf. Comp. Vis.*, pp. 11975–11986, 2023.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-CLIP: Unlocking the long-text capability of CLIP. In *Proc. Eur. Conf. Comp. Vis.*, pp. 310–325, 2024.

Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. DreamLIP: Language-image pre-training with long captions. In *Proc. Eur. Conf. Comp. Vis.*, pp. 73–90, 2024.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based language-image pretraining. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 16793–16803, 2022.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *Int. J. Comput. Vision*, 127(3): 302–321, 2019.

Fatimah Zohra, Chen Zhao, Hani Itani, and Bernard Ghanem. β-CLIP: Text-conditioned contrastive learning for multi-granular vision-language alignment. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 680–689, 2026.

A Data Curation Pipeline

We randomly sample images from CC3M (Sharma et al., 2018) and CC12M (Changpinyo et al., 2021) to construct the MG-Data dataset.

A.1 Image-Aware Curation

We begin by producing long and detailed captions. Each image is passed to the Qwen-VL-Max API with a long-caption generation prompt that emphasizes semantic completeness, precision, and linguistic fluency (see Figure 4). This step yields semantically rich, context-aware descriptions that thoroughly capture the visual content of each scene.

A.2 Object-Aware Curation

At this stage, we construct object-aware annotations consisting of fine-grained attribute descriptions for each object, paired with masks that precisely denote their spatial extent. We first feed the long captions into the Qwen-Max API using an object-extraction prompt (see Figure 5), which returns a list of salient object descriptions enriched with concrete attributes. Each extracted object caption is then provided to MLCD-Seg (An et al., 2024; Wu et al., 2025) with the prompt "Could you provide a segmentation mask for {object caption} in this image?", producing a mask that localizes the described object.

A.3 Region-Aware Curation

Finally, we identify groups of semantically related objects by feeding the image captions and their corresponding object lists into the Qwen-Max API (see Figure 6). The model extracts region-level captions describing the shared context or interactions among the selected objects. Each region caption is paired with its associated object set, and the corresponding region mask is constructed by merging the masks of all involved objects. The resulting annotation links each multi-object description to visual support covering the objects mentioned in the caption.

A.4 Post-processing

In the post-processing stage, we filter out low-quality object masks to improve the quality of object-aware and region-aware annotations. We begin by discarding masks whose area exceeds 85% of the entire image, as such oversized masks typically indicate segmentation failures rather than meaningful object boundaries. For any pair of masks with an IoU greater than 0.9, we randomly remove one to avoid redundant duplicates. Furthermore, if two masks have an IoU above 0.8 and the smaller mask contains more than ten connected components suggesting substantial noise or fragmentation, we discard the smaller mask.

To further improve mask quality, we apply a morphological opening operation to smooth jagged edges and remove small spurious structures, yielding cleaner and more stable object boundaries. Because patch pooling requires at least one visual patch satisfying $\alpha_{u,v} \geq \alpha_{\mathrm{adp}}$, we further remove object annotations whose masks yield an empty selected patch set, i.e., $\mathcal{S}(M) = \varnothing$. After filtering, we generate region-level descriptions and their associated object sets using only the retained object annotations. Each region mask is constructed as the union of the masks of its associated objects. Overall, this process removes approximately 20% of initially generated masks while enhancing the consistency and precision of region-level annotations.

To prevent evaluation contamination, we screened all 6M MG-Data records against the 100K unique evaluation images used in our experiments. We used SHA-256 hashes of the original image bytes and decoded pixel arrays for exact matching, and pHash for near-duplicate retrieval, followed by pixel-similarity analysis, SIFT-RANSAC matching, and manual verification. This process identified 90 high-confidence matching pairs corresponding to 87 unique MG-Data records, all of which were removed from the final training set. Examples of curated data are shown in Figures 7–9.

B More Experiments

B.1 Comparison with Triplet Loss

To further evaluate LPO, we compare it with a standard triplet loss that directly encourages a larger similarity gap between matched and hard-negative pairs using long captions. In contrast, LPO treats short-caption similarities as stop-gradient references and encourages the long-over-short similarity gain to be larger for matched pairs than for mined hard negatives. As shown in Table 6, LPO performs better than or comparably to the triplet loss on both CLIP and SigLIP 2. These results suggest that encouraging larger long-over-short similarity gains for matched pairs than for hard negatives benefits long-caption discrimination beyond directly optimizing their similarity separation.

Table 6: Comparison of LPO with a standard triplet loss on CLIP and SigLIP 2. All variants retain the same intra- and cross-granularity alignment objectives and differ only in the additional global discrimination constraint. We report I2T and T2I Recall@1 for image–text retrieval, Top-1 accuracy on ImageNet-v2, and mIoU on VOC.

Backbone	Constraint	MSCOCO		Flickr30K		DCI		Urban1K		IN-v2	VOC
		I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I		
CLIP	$\mathcal{L}_{\text{triplet}}$	60.5	41.6	89.0	73.5	57.8	58.3	92.3	91.8	61.3	66.0
	$\mathcal{L}_{\text{lpo}}$	60.9	41.8	89.4	73.6	58.4	59.1	93.0	92.1	61.8	66.0
SigLIP 2	$\mathcal{L}_{\text{triplet}}$	71.2	54.7	95.4	82.6	58.0	58.1	87.8	86.4	72.7	77.7
	$\mathcal{L}_{\text{lpo}}$	72.1	54.8	95.4	83.3	58.6	58.9	88.2	87.3	72.8	78.1

B.2 Sensitivity to the Number of Hard Negatives

We evaluate LPO with $K \in \{2, 5, 10\}$ while keeping all other training settings fixed. As shown in Table 7, performance remains stable across retrieval, classification, and segmentation on both CLIP and SigLIP 2. For each backbone, the difference between the highest and lowest values of each metric across the tested K values is at most 0.5 percentage points. Increasing K provides no consistent benefit, suggesting that using a small set of the mined hardest negatives is sufficient under the evaluated settings. We therefore use $K = 2$ as the default.

Table 7: Sensitivity of MaskAlign to the number of hard negatives K in LPO on CLIP and SigLIP 2, with all other training settings fixed.

Backbone	K	DCI		Urban1K		MSCOCO		Flickr30K		ImageNet-1K	COCO BBox Cls.		VOC Seg.
		I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I		Top-1	Top-5	
CLIP	2	58.4	59.1	93.0	92.1	60.9	41.8	89.4	73.6	68.1	57.6	85.3	66.0
	5	58.5	59.0	92.8	92.0	61.1	41.9	89.3	73.5	67.9	57.5	85.3	66.0
	10	58.0	59.1	92.8	91.9	61.0	41.9	89.3	73.7	67.9	57.6	85.3	66.0
SigLIP 2	2	58.6	58.9	88.2	87.3	72.1	54.8	95.4	83.3	79.6	39.4	82.4	78.1
	5	58.5	58.4	88.4	87.1	71.9	54.9	95.2	83.0	79.6	39.2	82.2	77.9
	10	58.5	58.7	88.4	87.2	71.7	54.9	95.1	83.2	79.6	39.3	82.2	78.0

B.3 Effect of Training Mask Perturbations

We examine how the spatial accuracy of training masks affects MaskAlign by comparing the original masks with two perturbed variants: masks enlarged by a linear factor of 1.2 and masks shifted in random directions with a bounding-box-normalized displacement of 0.2. Shifting preserves mask shape and scale, and both variants are subsequently clipped to the image boundaries. We train CLIP- and SigLIP 2-based MaskAlign under each mask setting, keeping the images, captions, other training settings, and evaluation protocol unchanged.

As shown in Table 8, the original masks consistently outperform both perturbed variants on all five segmentation benchmarks and COCO bounding-box classification across both backbones. The degradation is particularly pronounced for shifted masks on SigLIP 2, with segmentation drops of 2.7–4.4 mIoU. These results support the importance of spatially accurate mask supervision, par-

Table 8: Effect of training mask perturbations on MaskAlign with CLIP and SigLIP 2, with all other training settings fixed.

Backbone	Training masks	Image-Level Tasks			BBox Classification		Segmentation				
		MSCOCO	DCI	IN-v2	COCO	LVIS	VOC	Con	COCO	City	ADE
CLIP	Original	60.9/41.8	58.4/59.1	61.8	57.6/85.3	22.7/45.0	66.0	25.2	16.1	13.7	12.3
	Enlarged	60.8/41.9	58.0/59.0	61.4	56.1/84.6	22.9/45.4	64.4	24.4	15.5	13.3	11.9
	Shifted	60.6/41.8	57.7/58.8	61.5	53.7/83.8	22.7/45.1	65.1	24.8	15.7	12.9	12.1
SigLIP 2	Original	72.1/54.8	58.6/58.9	72.8	39.4/82.4	23.9/49.1	78.1	30.0	19.8	20.2	15.5
	Enlarged	71.6/54.7	58.5/58.3	72.8	38.5/79.6	23.8/48.6	77.1	28.6	18.5	18.7	14.6
	Shifted	71.4/54.6	58.5/58.7	72.7	36.5/75.3	22.6/46.9	74.9	25.8	17.1	15.8	12.7

ticularly for dense prediction. The comparatively smaller changes in global retrieval suggest that image-level retrieval alone may not fully reflect the spatial fidelity of learned local representations.

B.4 Qualitative Analysis of LPO

To illustrate the effect of LPO on long-caption retrieval, we visualize selected T2I retrieval cases, comparing the top-1 results of MaskAlign trained with and without LPO. For readability, the figures summarize key details from each long caption, while the full original captions are used as retrieval queries.

Retrieval rescue cases Figure 10 presents four examples in which adding LPO raises the ground-truth image from rank 2 or 3 to rank 1. In each case, the variant without LPO retrieves an image that matches the broad scene or object category but misses distinguishing details in the caption. The aircraft example specifies a gray MiG-31 with an identification sign and other aircraft nearby. The boat example describes two long boats with differently dressed rowers and ceremonial figures. The dress example includes distinctive motifs and hem patterns, while the street example specifies fish-shaped windsocks extending between buildings. The variant with LPO retrieves the GT images that match these combinations of details.

These cases illustrate how details beyond the shared scene or object category can distinguish otherwise similar candidates. The observed ranking changes are consistent with LPO's objective of encouraging larger long-over-short similarity gains for matched pairs than for hard negatives. Together with the quantitative results, these examples support the effectiveness of LPO in fine-grained image–text matching.

Retrieval failure cases Figure 11 presents selected failure cases in which MaskAlign with LPO does not rank the ground-truth image first. These cases involve closely related candidates that match many of the same details in the captions. Distinguishing these images requires resolving subtle differences in specific attributes and contextual details, which remains challenging even with LPO.

System Message:

You are a professional image captioner. Please describe the image in a natural, continuous paragraph. Here are some requirements:
1. Capture **all observable elements** with precision.
2. **Label each object with a concrete and distinguishing expression of its attributes** (e.g "a woman in a red dress", "a woman in a black jacket", "a steel chair", "a wooden chair", etc, avoid expressions like "a woman" or "a chair"). One object's expression should always remain the same throughout the description.
3. The description should seamlessly integrate the scene, setting, atmosphere, objects, objects' intrinsic attributes (color, texture, size), spatial relationships, interactions, actions, and depth.
4. Ensure coherence with natural transitions, avoiding a list-like structure while maintaining precision and completeness.
5. Only mention what you are certain about, and **avoid** descriptions that may be **ambiguous or inaccurate**.

User Message:

According to the requirements above, please generate the caption for the image.



Figure 4: Prompt used for long-caption generation with Qwen-VL-Max in the curation of image-aware annotations.

System Message:

You are an AI assistant skilled in information extraction and linguistic analysis. Your task is to extract the main objects from the given image caption. Follow these guidelines carefully:
1.Focus on the primary objects in the scene, including key attributes that are essential to identifying the object. If the object has significant attributes or parts described together, treat them as a single entity.
2.When extracting objects, avoid splitting objects that are described together (e.g., if an object is described with attributes or multiple parts, treat it as a single entity unless those parts are independently significant)
3.The extracted objects should be distinct physical entities. Do not include abstract descriptions that do not refer to separate objects.
4.If the caption describes a broad background area that physically exists in the scene (e.g., the sky, a grassy field, the ocean, a forest backdrop), and it contributes meaningfully to the scene, treat it as a separate object and extract it as well.
5.Reflect several times to ensure that all the essential objects are identified, but avoid redundancy. Ensure that only the major objects that physically exist in the scene are listed.
6.Output only a Python list of the extracted objects, without any extra text.

For example, the image caption is "The image contains an expensive chair. The chair has black armrests, peach-colored legs, and is placed next to the dining table.", the extracted object list should be a Python list: ["a chair with black armrests and peach-colored legs", "dining table"]

User Message:

Here is the image caption: {caption}. According to the requirements above, please extract the object list.

Figure 5: Prompt used for object-extraction with Qwen-Max in the curation of object-aware annotations.

System Message:

You are an AI assistant skilled in object relationship analysis and region-specific caption generation. Your task is as follows:

Given:
- An **image caption** that describes the overall scene.
- A list of **extracted objects**, each extracted from the caption and associated with a corresponding object mask.

Please perform the following steps:

1. **Relationship Identification**:
- Analyze the image caption and identify **meaningful relationships involving two or more objects** from the object list.
- For each relationship, extract the segment from the caption that **best describes** the interaction or shared context of these objects.
- Do not include any irrelevant background details or positional information (e.g., "to the left of the building", "in the background", etc.). Focus only on the relationship between the objects.
- **Do not omit** any multi-object interaction. Ensure **all relevant objects** from the object list are included.

2. **Region Caption Generation**:
- For each group of related objects, generate a region-specific caption that focuses **only on the mentioned objects**.
- You need to **rephrase or shorten** the original caption segment, and must **preserve the original meaning** and ensure the result is **fluent and self-contained**.
- The caption should **exclude any elements not directly related to** the listed objects, even if they appear in the original caption.
- Since only the masks of the related objects are visible, **do not include background descriptions or other unrelated objects**.

3. **Output Format**:
- Return a Python list of dictionaries in the following format:
```
[{"caption": "Caption of region_1 in the image.",
  "related_objects": ["object_1", "object_2", ...]},
...]
```
- Do **not** output any explanation, commentary, or additional text outside this list.

Example:
If the object list includes "a large rock" and "moss", and the original caption says: "Behind the tree, a large rock is partially covered by moss, giving it an ancient, weathered look.", the output should be:
```
[{"caption": "A large rock is partially covered by moss, giving it an ancient, weathered look.",
  "related_objects": ["a large rock", "moss"]}
]
```

User Message:

- Here is the image caption: {**caption**}.
- Here is the object list: {**objects_list**}.
- Return the json format.

Figure 6: Prompt used for region-caption extraction with Qwen-Max in the curation of region-aware annotations.

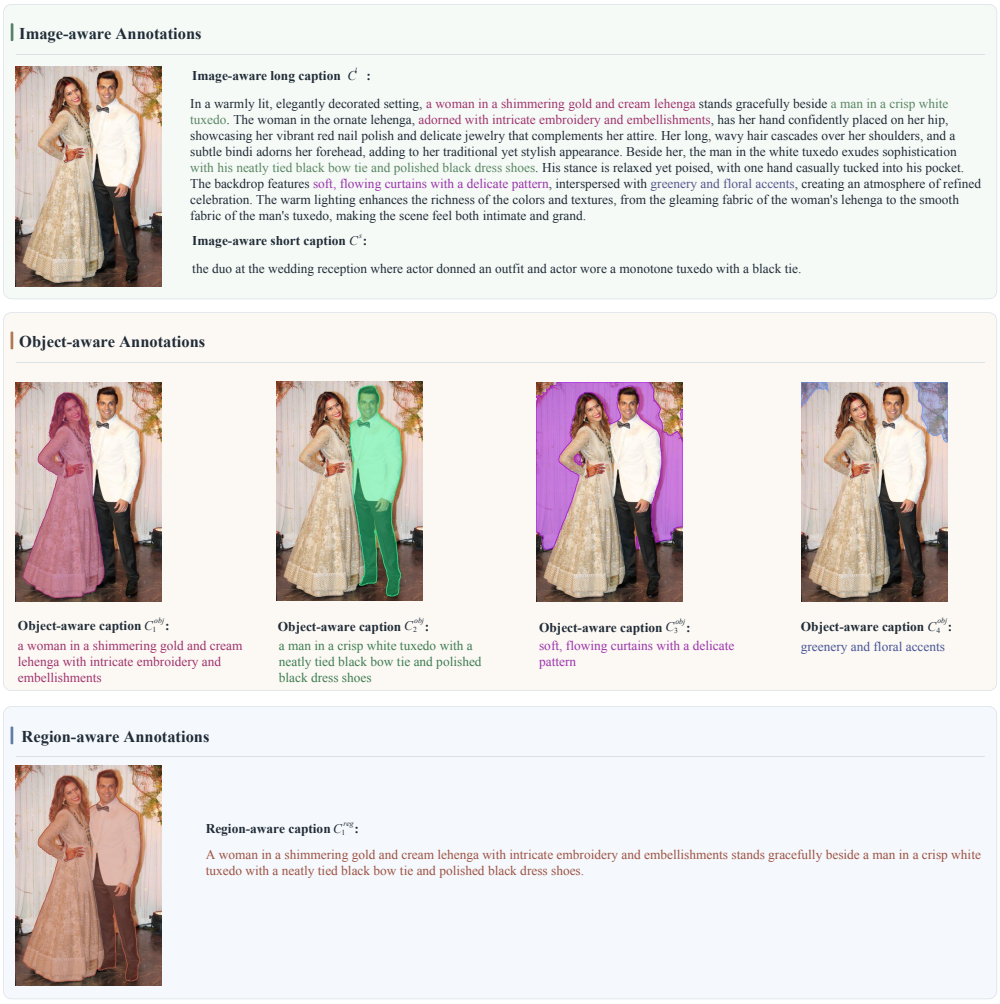

Figure 7: One sample from MG-Data. Each example includes image-aware annotations (long and short captions), object-aware annotations (object masks and captions), and region-aware annotations (region masks and descriptions).

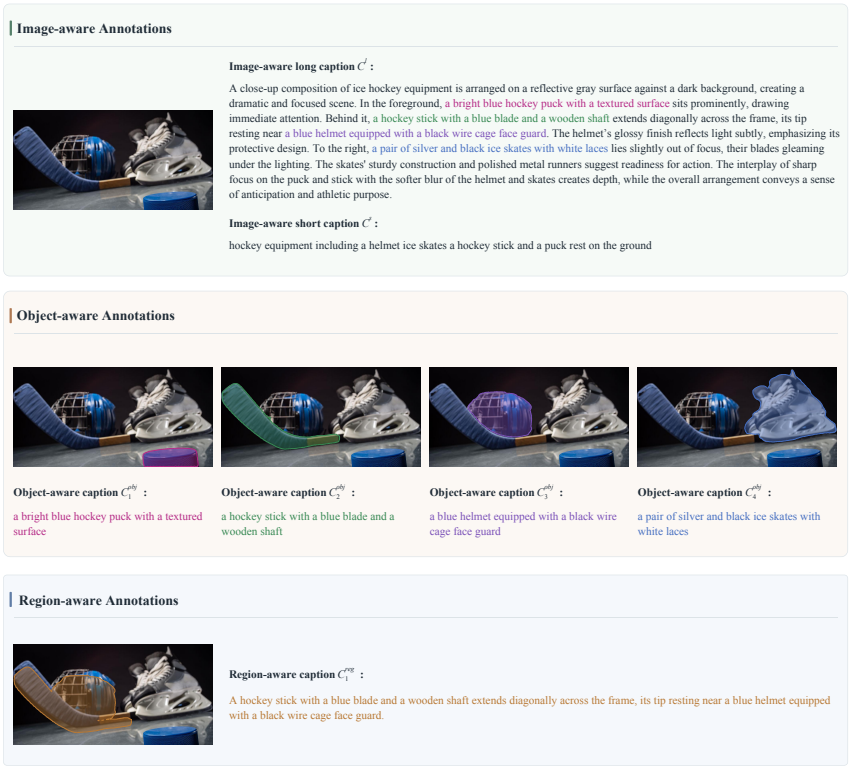

Figure 8: Additional MG-Data samples.

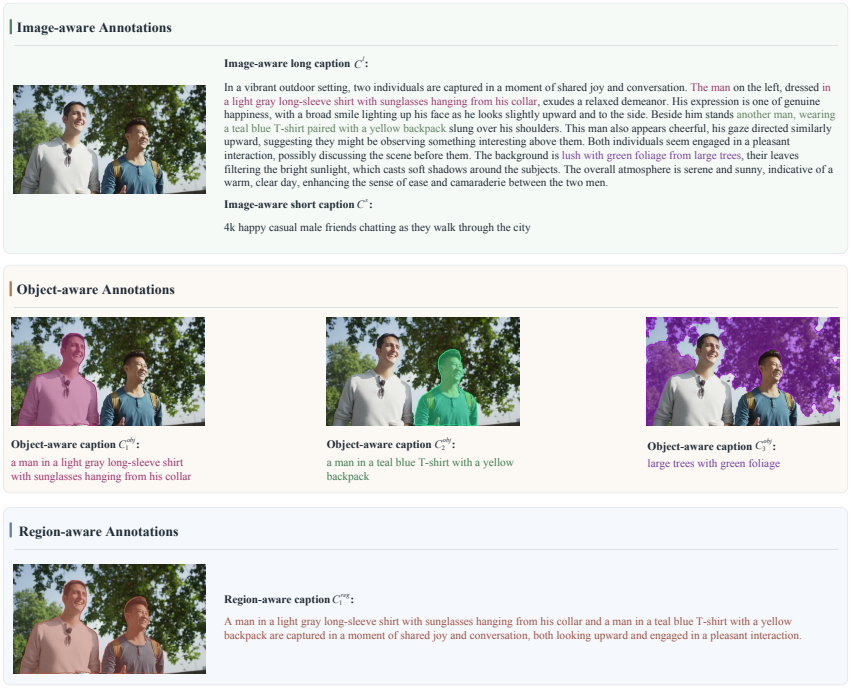

Figure 9: Additional MG-Data samples.

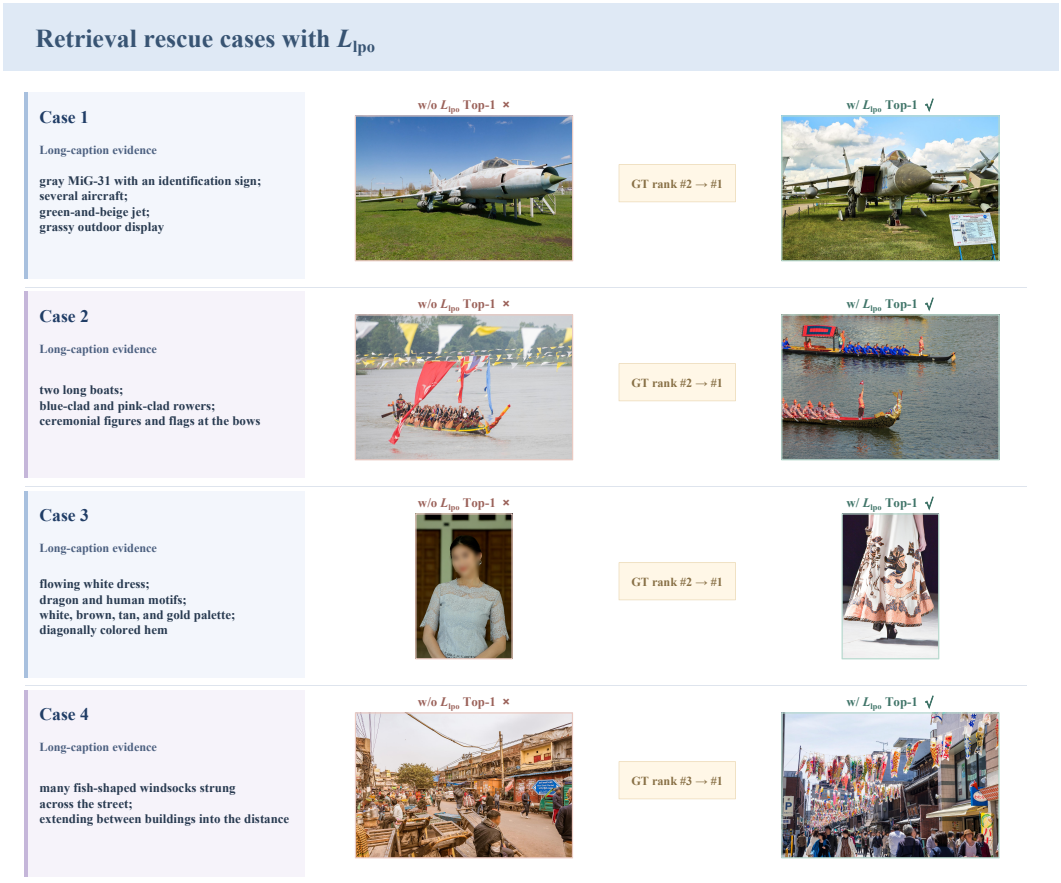

Figure 10: Retrieval rescue cases with LPO.

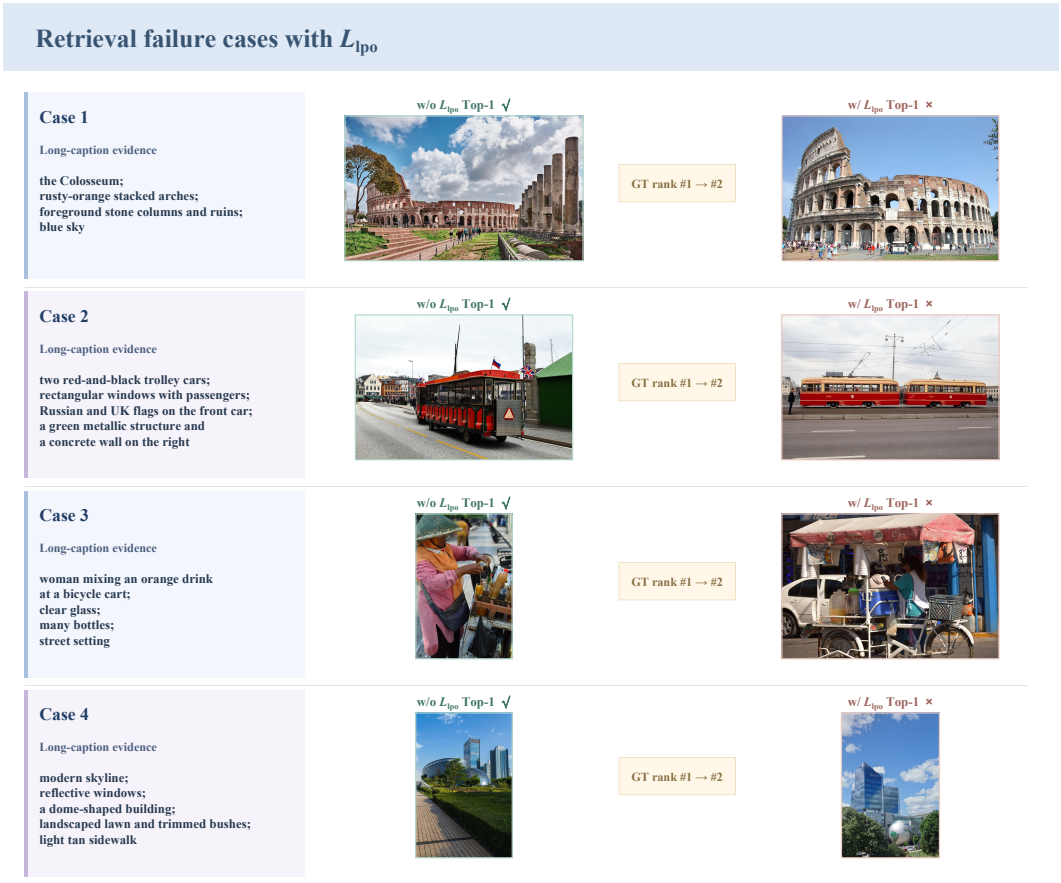

Figure 11: Retrieval failure cases with LPO.