LARE: Low-Attention Region Encoding for Text–Image Retrieval

Anonymous CVPR submission

Paper ID

Abstract

*Image retrieval in crowded scenes is particularly challenging due to the salience bias of conventional visual encoders, which tend to focus on dominant objects while neglecting low-attention regions that are often crucial for fine-grained retrieval. We propose **LARE** (Low-Attention Region Encoding), a framework that explicitly models these overlooked regions. LARE adopts a dual-encoding strategy that encodes low-attention regions of an image and the full image in parallel, leading to more diverse and informative image embeddings. To evaluate image retrieval performance in challenging crowded scenes, we introduce **Dense-Set**, a challenging subset derived from COCO and Flickr30K. In this subset, images are re-captioned to provide richer descriptions of low-attention or previously overlooked regions. This dataset highlights the limitations of existing retrieval models and enables a more rigorous evaluation under densely crowded scene conditions. Experimental results demonstrate that the proposed framework improves retrieval performance by preserving subtle, non-dominant visual cues within the shared latent space.*

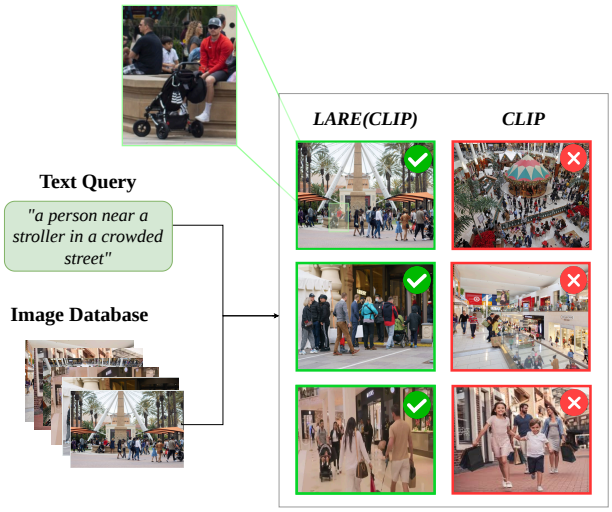


Figure 1. **Fine-grained retrieval in dense scenes.** For the query *"a person near a stroller in a crowded street"*, LARE retrieves results that preserve the stroller-related local cue, while CLIP tends to favor globally similar crowded scenes. Green checks indicate relevant matches; red crosses indicate mismatches.

1. Introduction

Text-to-image retrieval retrieves images from large collections that best match a natural-language query. This capability is central to many real-world applications, including multimedia search engines, content recommendation systems, digital asset management, and large-scale visual indexing for web platforms. More broadly, cross-modal retrieval enables intuitive natural-language interaction with visual data and has become a key component in modern multimodal AI systems. [3, 7, 8, 10, 12, 15, 16, 18, 22].

Recent advances in large-scale vision–language pretraining have significantly improved cross-modal retrieval by learning shared embedding spaces in which images and text can be compared directly. Contrastive models such as CLIP [18] and ALIGN [10] learn aligned visual and textual representations using massive image–text datasets, enabling strong zero-shot transfer across many tasks without task-specific training. In these models, an image encoder and a text encoder project inputs from each modality into a common embedding space, and retrieval is performed by ranking the similarity between their representations. This paradigm has become the dominant approach for cross-modal retrieval and underlies many modern multimodal systems [5, 9–11, 13, 18, 24].

Despite their success, current vision-language encoders mainly rely on a *global image embedding* that summarizes the entire image into a single representation. Although effective for many queries, this representation often emphasizes the most visually salient objects or scene context while underrepresenting smaller or less prominent elements. As a result, retrieval models may overlook visually relevant cues that occupy only a small portion of the image. This limitation is particularly evident in dense scenes with many objects, where correct retrieval may depend on attributes or objects that are not dominant in the global representation.

Previous work has shown that vision-language models can struggle to localize fine-grained visual evidence and often prioritize coarse scene semantics over detailed object-level information [21].

In this work, we address this limitation by recovering information from image regions that receive little attention in the global representation. Our key observation is that transformer-based vision encoders implicitly encode spatial attention signals that reveal which regions contribute less to the final embedding. Rather than relying solely on the global representation, we exploit these signals to identify under-attended regions that may contain discriminative visual cues relevant to the query.

We propose Low-Attention Region Encoding (LARE), a training-free framework that augments standard dual-encoder retrieval models with region-level evidence. Given an input image, LARE extracts low-attention regions from the encoder's attention maps and re-encodes them to complement the global image embedding. During retrieval, the similarity between the text query and both global and regional representations is evaluated using a confidence-gated scoring mechanism.

To evaluate retrieval under challenging conditions, we introduce Dense-Set, a curated subset of COCO [14] and Flickr30K [23] that emphasizes crowded scenes and rare objects. The dataset contains images with many detected objects and at least one rare object instance, along with re-captioned descriptions that highlight these underrepresented elements.

Experiments show that LARE consistently improves retrieval performance in dense scenes while preserving the ranking behavior of the original encoder on standard benchmarks, without requiring additional training, parameters, or architectural modifications.

Our contributions can be summarized as follows:

- We propose **LARE**, a training-free retrieval framework that augments global image embeddings with region-level representations extracted from low-attention areas.
- We introduce **Dense-Set**, a curated benchmark designed to evaluate retrieval performance in crowded scenes containing rare or visually subordinate objects.
- We conduct extensive experiments and ablation studies demonstrating consistent improvements on dense retrieval benchmarks across multiple backbone encoders while preserving performance on standard datasets.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 introduces the Dense-Set and its construction pipeline. Section 4 presents the proposed LARE retrieval framework. Section 5 reports experimental results and analysis on both standard benchmarks and Dense-Set. Finally, Section 6 concludes the paper.

2. Related Work

This work is related to research on text-to-image retrieval using vision–language models, methods for fine-grained image–text alignment, and approaches to retrieval in dense, visually complex scenes.

2.1. Text-to-Image Retrieval

Text-to-image retrieval aims to retrieve images that match a natural language query, and it is a fundamental task in vision–language understanding [5, 9–11, 13, 18, 24]. Early approaches learned joint embedding spaces using convolutional neural networks for visual encoding and recurrent networks for text representation [6, 19]. More recently, large-scale vision–language pretraining has significantly improved retrieval performance by leveraging massive collections of image–text pairs [13, 18, 25].

Dual-encoder architectures have become the dominant paradigm for this task. Models such as CLIP and ALIGN learn aligned image and text representations using contrastive learning over large-scale datasets, enabling strong zero-shot retrieval performance across multiple benchmarks [10, 18]. In these models, the image and text encoders independently project each modality into a shared embedding space, allowing efficient similarity computation and scalable retrieval. Subsequent works have further improved representation quality and training efficiency. For example, BLIP introduces bootstrapped caption generation to enhance multimodal representation learning [13], while SigLIP replaces the traditional softmax contrastive loss with a sigmoid loss to improve scalability and training stability [24].

Despite their strong performance, dual-encoder retrieval models typically rely on a *global image embedding* that summarizes the entire image into a single vector. While effective for many queries, such representations may underrepresent localized visual evidence when relevant objects occupy small or visually subordinate regions within the image.

2.2. Fine-Grained Vision–Language Alignment

To address the limitations of global representations, several works explore fine-grained alignment between image regions and textual tokens. FILIP introduces a late-interaction mechanism that computes token-level similarity between image patches and textual tokens, enabling finer-grained cross-modal alignment while maintaining efficient inference [22]. PyramidCLIP further improves alignment by introducing hierarchical feature representations that capture visual semantics at multiple levels of granularity [7].

Another line of work focuses on region-level representations. RegionCLIP extends contrastive language-image pretraining to region-based representations, enabling alignment between **t**extual concepts and localized image re-

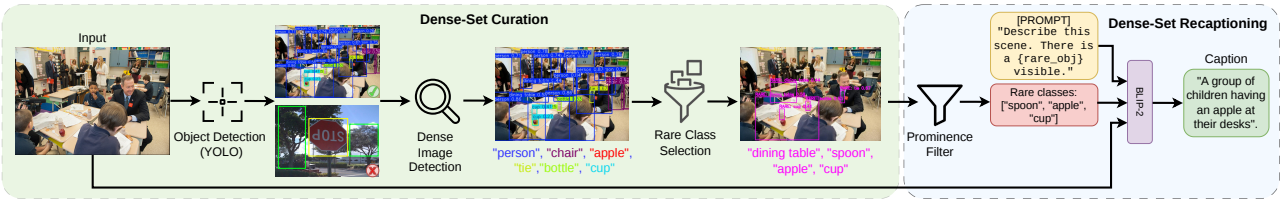


Figure 2. Dense-Set curation pipeline. We first detect objects with YOLO and rank images by total object count, retaining the top 10% as the *High-Density Subset* (dense candidate pool). We then apply rare-class filtering and keep images containing at least one single-instance class to form the final Dense-Set.

gions [26]. More recently, methods such as ELIP introduce lightweight text-guided visual prompts that condition the image encoder on the query, improving retrieval performance without retraining large backbone models [25].

While these approaches improve fine-grained alignment, many require additional training, architectural modifications, or query-conditioned representations, thereby increasing computational complexity. Unlike prior approaches that require retraining or query-conditioned encoders, our method augments global representations with region-level embeddings extracted at inference time, thereby improving retrieval in dense scenes while preserving the efficiency of dual-encoder architectures.

2.3. Retrieval in Dense and Complex Scenes

Text-to-image retrieval becomes particularly challenging in crowded scenes and long-tail object distributions, where relevant evidence may correspond to small or rare objects. Datasets such as COCO and Flickr30K contain complex scenes with multiple objects, occlusions, and visual clutter, making global image representations insufficient for capturing fine-grained attributes [14, 17]. In such scenarios, correct retrieval may depend on localized visual cues that are not dominant within the scene. To address this, prior work has explored combining global and local representations, for example by leveraging local features to refine global similarity rankings [1].

Recent studies have also shown that attention maps produced by vision transformers encode implicit spatial signals that indicate which regions contribute most to the final representation. These signals have been used for interpretability and weak localization tasks, revealing how visual transformers allocate attention across spatial regions. Concurrent work explores a related inverse-attention idea for video retrieval [2], fusing regional and global scores via a hard maximum; in contrast, LARE targets image retrieval and introduces confidence-gated fusion together with the curated Dense-Set benchmark.

Motivated by these observations, our work leverages the internal attention structure of vision transformers to identify *low-attention regions* that may contain underrepresented visual evidence.

3. Dense-Set Dataset

To evaluate the proposed methodology, we construct **Dense-Set**, a curated benchmark of visually dense scenes. The goal is to create a challenging evaluation subset containing crowded images with multiple object instances and underrepresented classes. To this end, we develop an automated pipeline, illustrated in Figure 2. In the following subsections, we describe the main stages of this pipeline.

3.1. Dense-Set Construction

This stage of the pipeline, illustrated in the first half of Figure 2, focuses on identifying densely populated images that contain underrepresented object instances. We begin by processing all images from the COCO [14] and Flickr30K [23] test splits using a YOLO object detector [4]. For each image, the detector outputs bounding boxes and class predictions, from which we compute three image-level statistics: (i) the total number of detected objects, (ii) the number of unique object categories, and (iii) per-class instance frequencies.

To construct the dense candidate pool, images are ranked in descending order by total object count, and the top **10%** are selected. This step favors crowded scenes with high object density and diverse visual content. Within this dense candidate set, we identify *rare classes* at the image level, defined as object categories that appear exactly once in a given image. In crowded scenes, such single-instance categories often correspond to small or low-salience objects that are easily overlooked by global representations.

The final Dense-Set subset consists of images that (1) belong to the dense candidate pool and (2) contain at least one rare-class instance. This selection strategy yields a benchmark with significantly higher object density and class diversity than the original splits, thereby creating a more challenging setting for fine-grained text-to-image retrieval.

Table 2 summarizes the three stages shown in Figure 2: the Original Test Set, the High-Density Subset (top 10% by object count), and the final Dense-Set after rare-class filtering. For each stage, we report the number of images, the average number of detected objects per image, and the

Table 1. Examples from Dense-Set with rewritten captions highlighting rare or low-attention objects for more challenging dense-scene evaluation.

Dataset	**COCO**	**COCO**	**Flickr30K**	**Flickr30K**
Image				
Original Caption	Car driving down a road behind a lot of sheep.	A cat lying down on a desk by a computer keyboard.	A group of men wearing sweaters are dining in a hall.	A crowd of people is standing outside next to a street.
Rare Class	Dog	Sports ball	Fork	Handbag
Rewritten Caption	A photo of a dog standing on the side of a road with a herd of sheep.	A sports ball sitting on top of a desk.	A fork placed in the middle of a group of men sitting at a table.	A handbag on the ground in front of a crowd of people.

Table 2. Stage-wise statistics of Dense-Set curation for COCO and Flickr30K

Dataset	**Split**	**# Images**	**Avg. Objects**	**Avg. # Classes**
COCO	Original Test Set	40,504	6.71	2.85
	High-Density Subset	4,050	21.63	4.82
	Dense-Set	3,089	21.63	5.47
Flickr30K	Original Test Set	31,783	6.73	2.48
	High-Density Subset	3,178	19.40	4.38
	Dense-Set	2,477	19.55	4.85

average number of object classes. The final curated Dense-Set contains images with substantially more objects and a broader set of object categories compared to the original splits. These characteristics make Dense-Set particularly suitable for evaluating retrieval models in visually dense environments, where important objects may appear in low-attention regions and are more likely to be overlooked by standard global representations.

3.2. Dense-Set Re-captioning

The second stage of the pipeline, illustrated in the second half of Figure 2, focuses on regenerating captions for the curated Dense-Set images. The goal of this re-captioning step is to produce more challenging textual descriptions that explicitly emphasize low-attention regions, i.e., rare-class instances. In contrast, the original dataset captions typically describe the dominant scene context and often overlook small or underrepresented objects. For each image in Dense-Set, we first filter rare-class detections whose bounding boxes occupy a large fraction of the image area (e.g., greater than 15%). Such instances are likely to correspond to visually dominant objects rather than genuinely low-salience elements. This filtering ensures that the captioning process focuses on secondary or background objects that are more likely to be ignored by global visual representations. The rare-class-filtered labels are then used as guidance for a vision-language model (BLIP-2). Specifically, we prompt the model to use class-aware templates *(e.g., "a photo of a [class]")* to encourage explicit mention of these underrepresented objects in the generated description. The model takes both the image and the guided prompt as input and outputs a single caption in the standard COCO format. By shifting the caption focus from general scene-level descriptions to fine-grained object-level details, this re-captioning process produces a more demanding evaluation setting for text-to-image retrieval in dense scenes.

Examples of the curated Dense-Set and their rewritten captions are shown in Table 1. For each image from COCO and Flickr30K, we identify a rare or low-attention class and rewrite the original caption to explicitly describe the overlooked object. This shifts the textual focus from general scene context to fine-grained object-level details, thereby making dense-scene retrieval evaluation more challenging.

4. Methodology

We introduce Low-Attention Region Encoding (LARE), a training-free framework that enhances visual semantic search by recovering information from regions typically underemphasized by standard vision encoders. Our approach follows a three-stage pipeline illustrated in Figure 3: (1) Low-Attention Region Detection, (2) Regional Encoding, and (3) Confidence-Gated Scoring.

4.1. Low-Attention Region Detection

The first stage identifies non-dominant visual cues by analyzing the internal self-attention signals of a frozen vision encoder. Given an input image I, we extract the self-attention tensor from an intermediate layer ℓ. For each head h, let $\mathbf{A}^{(h)} \in \mathbb{R}^{HW \times HW}$ denote the patch-to-patch attention matrix.

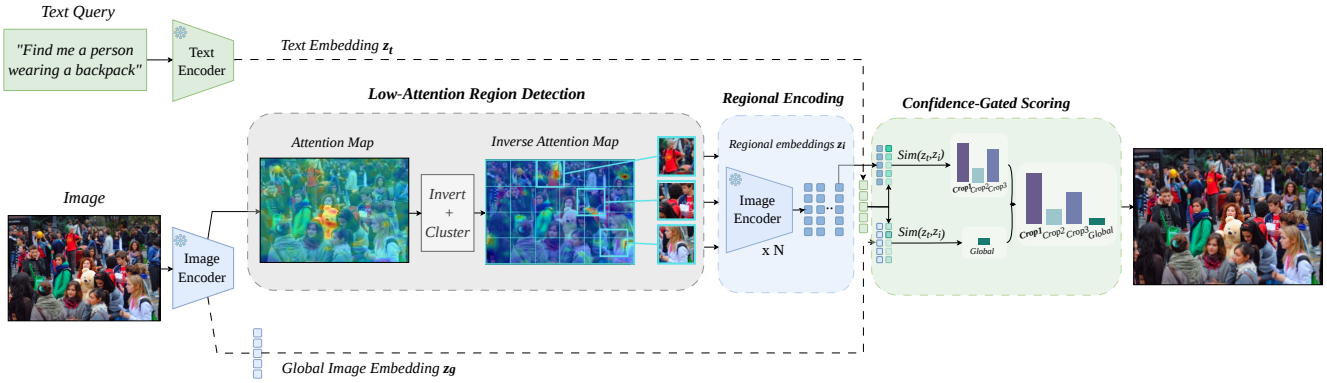


Figure 3. LARE pipeline: A single forward pass produces both a global image embedding and a spatial attention map. Inverting the attention map highlights under-attended regions, which are clustered into candidate crops and then re-encoded independently. A confidence gate determines whether regional evidence should be used to adjust the final retrieval score.

We quantify the amount of attention each patch i receives from all other patches by calculating the column-wise sum:

$$a_i^{(h)} = \sum_j A_{j,i}^{(h)}, \qquad i \in \{1, \ldots, HW\} \tag{1}$$

Each map $a^{(h)}$ is reshaped to a spatial grid, min-max normalized, and averaged across the top-k heads (selected by spatial variance) to form a mean attention map $\bar{\mathbf{A}}$. We then derive an inverse-attention map:

$$\mathbf{M} = \mathbf{1} - \bar{\mathbf{A}} \tag{2}$$

where high values in $\mathbf{M}$ highlight patches that consistently receive minimal attention. We apply a sliding window and non-maximum suppression (NMS) on $\mathbf{M}$ to generate a set of N candidate regions, $\mathcal{R} = \{r_1, \ldots, r_N\}$. We analyze sensitivity to N in Appendix A.1, Figure 1.

4.2. Regional Encoding

The second stage encodes the image regions generated in the previous stage.

$$\mathbf{z}_i = f_v(r_i), \quad i = 1, \ldots, N \tag{3}$$

This produces a set of regional feature vectors $\{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$. Because the encoder weights are shared, these regional embeddings reside in the same feature space as the global representation, allowing for direct comparison with text embeddings without additional training.

4.3. Confidence-Gated Scoring

Finally, we integrate the global and regional information to compute a comprehensive retrieval score. While prior work fuses regional and global signals via a hard maximum [2], this can amplify spurious regional matches when the global embedding is already well-aligned. We instead introduce a confidence-gated fusion that defers to the global score when the model is confident, and only blends in regional evidence otherwise. First, we obtain the global image embedding $\mathbf{z}_g = f_v(I)$ and the text query embedding $\mathbf{z}_t = f_t(T)$. We define the global similarity as $s_g = \text{sim}(\mathbf{z}_t, \mathbf{z}_g)$ and the strongest regional match as $s_r = \max_i \text{sim}(\mathbf{z}_t, \mathbf{z}_i)$. To ensure robustness against regional noise, we gate the contribution of the regions based on the model's confidence in the global match. If s_g exceeds a confidence threshold τ, the final score remains $S = s_g$. If $s_g < \tau$ and a region outperforms the global match ($s_r > s_g$), we interpolate toward the regional score:

$$\alpha = \min\big(2(s_r - s_g), 0.5\big), \qquad S = (1-\alpha)s_g + \alpha s_r \tag{4}$$

where $\tau = 0.25$. We analyze the sensitivity to τ in Appendix A.1, Figure 1. This fusion logic ensures that regional evidence effectively "rescues" the ranking when the global embedding is insufficient, particularly in dense scenes targeting non-salient objects.

5. Results and Analysis

We evaluate LARE in a zero-shot image retrieval setting, where no additional training or fine-tuning is performed on the target benchmarks. Given a textual query, the task is to retrieve the most semantically aligned image from a candidate set. We compare the performance of LARE against several state-of-the-art vision–language retrieval models, including CLIP [18], SigLIP [24], and SigLIP 2 [20]. Evaluation is conducted on COCO [14] and Flickr30K [23], as well as their Dense-Set variants designed to emphasize crowded scenes and rare objects. Performance is reported using Recall@K metrics (R@1, R@5, R@10).

5.1. Zero-Shot Retrieval Results

Performance on standard datasets: As shown in Table 3, the first two column groups (COCO and Flickr30K)

Table 3. Zero-shot retrieval performance of baseline models and LARE pipeline on COCO and Flickr30K, along with their Dense-Set variants.

Model	ViT	COCO			Flickr30K			COCO-Dense			Flickr30K-Dense		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP [18]	L/14	36.10	61.10	71.44	65.00	88.00	92.62	17.79	35.85	45.11	3.48	11.97	16.33
SigLIP [24]	So/14	54.24	76.78	84.21	82.94	96.08	98.00	26.61	46.31	55.22	5.05	15.50	20.96
SigLIP 2 [20]	So/16	56.55	78.75	85.95	83.72	96.34	98.32	27.56	47.56	56.73	5.12	16.47	21.80
LARE (CLIP)	L/14	36.10	61.10	71.44	65.00	88.00	92.62	22.97	42.10	52.03	9.73	16.63	20.40
LARE (SigLIP)	So/14	54.26	76.80	84.24	82.94	96.12	98.00	29.94	50.17	59.26	12.33	19.87	24.10
LARE (SigLIP 2)	So/16	**56.56**	**78.78**	**85.97**	**83.76**	**96.38**	**98.34**	**31.00**	**51.45**	**60.67**	**13.28**	**21.11**	**25.10**

report results on standard benchmark splits. On these datasets, LARE maintains performance comparable to the underlying backbone models, with differences being marginal across all Recall@K metrics. In some cases, slight improvements are observed (e.g., +0.01 to +0.04 absolute R@1), but overall performance remains statistically similar. This indicates that the proposed method preserves the ranking behavior of the original encoders in conventional retrieval settings where global image representations are already sufficient.

Performance on Dense-Set: In contrast, the last two columns of Table 3 (COCO-Dense and Flickr30K-Dense) demonstrate substantial gains on the curated Dense-Set benchmarks. On COCO-Dense, LARE improves R@1 by +5.18 points (29% relative improvement) for CLIP, +3.33 points (12.5%) for SigLIP, and +3.44 points (12.5%) for SigLIP 2. On Flickr30K-Dense, the gains are even more pronounced: +6.25 points (180% relative improvement) for CLIP, +7.28 points (144% relative improvement) for SigLIP, and +8.16 points (159% relative improvement) for SigLIP 2.

These results show that while LARE preserves performance on standard benchmarks, it delivers large and consistent improvements in dense-scene retrieval scenarios, particularly where relevant objects are rare, small, or visually subordinate.

Cross-Backbone Generalization: The consistent improvement across diverse architectures (from CLIP to SigLIP 2) demonstrates that LARE operates as a general, plug-and-play inference refinement. It complements even the strongest modern encoders, suggesting that "salience bias" is a fundamental characteristic of global embeddings that persists despite scaling.

5.2. Qualitative Results

Figure 4 presents qualitative comparisons between the baseline encoder (SigLIP) and LARE on dense retrieval queries from COCO-Dense (Columns 1–2) and Flickr30K-Dense (Columns 3–4). For each query, the top-5 retrieved images are shown, and the ground-truth image is highlighted with a dashed box.

In the first example (COCO-Dense), the query *"A cyclist wearing a backpack next to a train station"* requires recognition of the backpack in addition to the cyclist and station context. The baseline ranks a generic cyclist at Rank 1, failing to capture the backpack attribute, while the correct image appears lower in the ranking. In contrast, LARE identifies the backpack as a localized discriminative cue and promotes the correct image to the top position for retrieval.

In the second example (Flickr30K-Dense), the query *"A person carrying a red bag in a busy outdoor market"* hinges on detecting the red bag within a crowded scene. The baseline retrieves general market scenes that align with the global context but miss the specific attribute described in the query. LARE successfully retrieves the image containing the person with the red bag at Rank 1, indicating improved alignment with fine-grained details.

These examples illustrate that improvements arise when relevant evidence is spatially localized and visually subordinate within the scene. By incorporating region-level representations, LARE resolves ambiguities that global embeddings alone fail to distinguish. When global similarity is already reliable, rankings remain unchanged, consistent with the confidence-gated design.

5.3. Inference Overhead

LARE introduces additional inference cost because each image is encoded once globally and up to N times for regional crops, where $N{=}5$ in our experiments. This computation occurs only when global similarity confidence falls below the threshold. No additional training, parameters, or model modifications are required. Because regional encoding is applied selectively, the overhead is bounded and can be adjusted through the crop count and confidence threshold.

In practice, this trade-off can be tuned to match deployment constraints. Lowering N or using a stricter confidence threshold reduces average latency, while higher N can improve recall on difficult dense-scene queries. This budget–

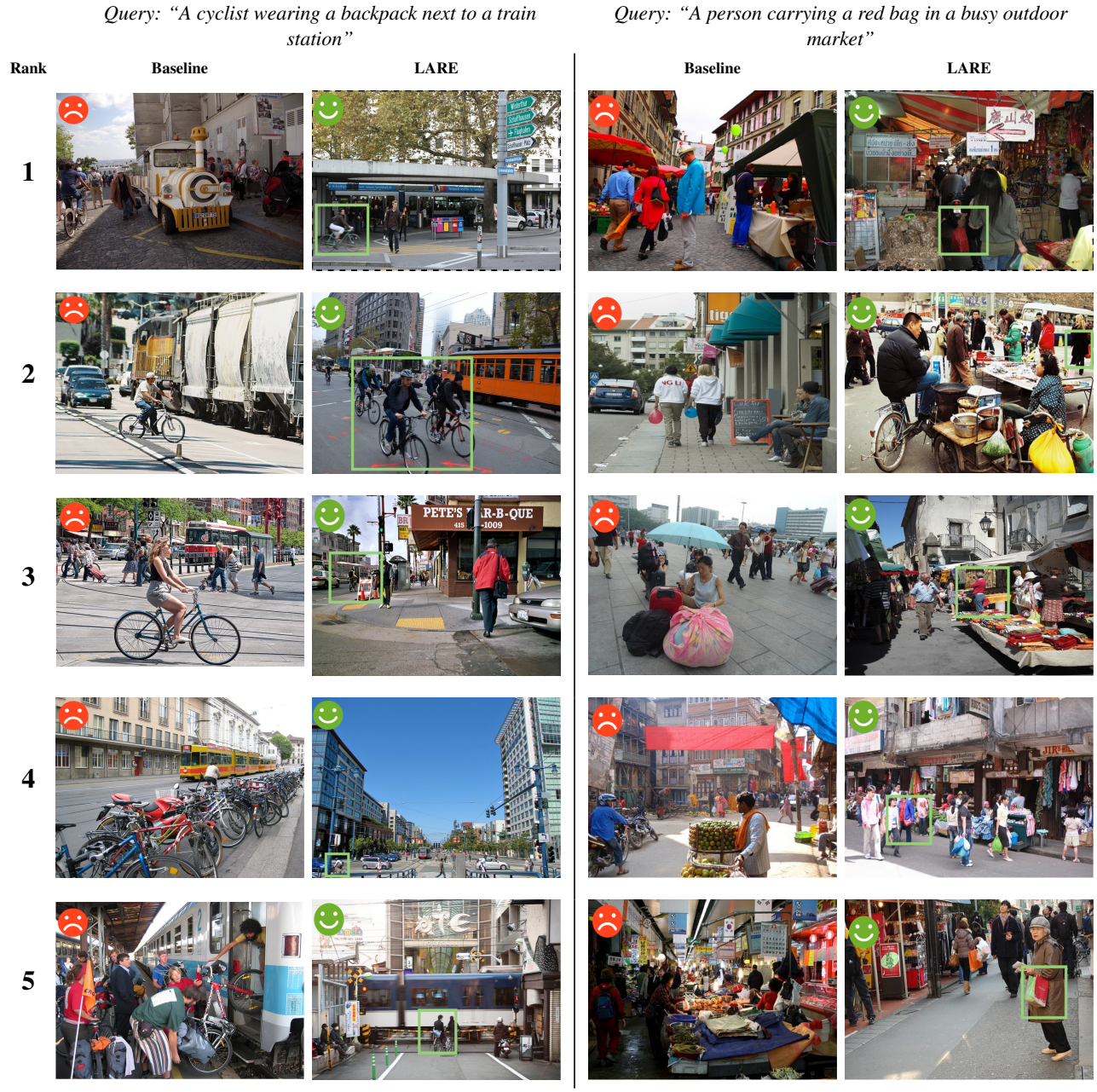


Figure 4. **Qualitative comparison between Baseline and LARE** on COCO-Dense (Cols. 1–2) and Flickr30K-Dense (Cols. 3–4). Top-5 retrieval results are shown; ground-truth is highlighted. LARE improves ranking by leveraging fine-grained, localized cues missed by the baseline.

accuracy control makes LARE suitable for staged retrieval pipelines, where a fast global pass is followed by selective regional refinement only for uncertain candidates. Hyperparameter ablations in the appendix further show that using fewer crops still yields robust gains, with performance saturating around $N{=}5$. A detailed latency analysis is left for future work.

6. Conclusion

We presented LARE, a training-free augmentation for text-to-image retrieval in crowded scenes. Our method mines low-attention regions from a frozen vision encoder, encodes these regions alongside the full image, and combines regional embeddings with the global image embedding at inference time. This simple test-time procedure improves

retrieval on Dense-Set variants that emphasize subtle and occluded content. We also introduced Dense-Set, a challenging crowded-scene benchmark derived from COCO and Flickr30K, where images are re-captioned to emphasize low attended areas. By shifting the focus toward fine-grained object, Dense-Set reveals the limitations of existing retrieval models and provides a more rigorous evaluation setting for densely crowded scenes.

For future work, we plan to make region selection more query-aware so that only the most informative crops are encoded, reducing compute while preserving accuracy gains. We also aim to strengthen fine-grained text–image alignment through patch-level interactions in the spirit of FILIP [22]. In addition, extending LARE to temporal retrieval settings is a promising next step, building on dual-encoder video retrieval formulations such as CLIP4Clip and Frozen in Time [3, 15].

References

[1] Dror Aiger, Bingyi Cao, Kaifeng Chen, and Andre Araujo. Global-to-local or local-to-global? enhancing image retrieval with efficient local search and effective global reranking. *arXiv preprint arXiv:2509.04351*, 2025. 3

[2] Faisal Aljehrai, Mohammed A. Alkhrashi, Alreem Almuhrij, Sarah Abuhimed, Noorh Aldossary, Abdullah Aldwyish, Raied Aljadaany, Huda Alamri, and Muhammad Kamran J. Khan. Look beyond saliency: Low-attention guided dual encoding for video semantic search, 2026. 3, 5

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1728–1738, 2021. 1, 8

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 3

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[6] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, Bejing, China, 2014. 2

[7] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 35:35959–35970, 2022. 1, 2

[8] Satya Krishna Gorti, Noel Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10562–10571, 2022. 1

[9] Yichen Huang, Yuchen Wang, and Yik-Cheung Tam. Uniter-based situated coreference resolution with rich multimodal input. *arXiv preprint arXiv:2112.03521*, 2021. 1, 2

[10] 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 *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1, 2

[11] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 1, 2

[12] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1

[13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1, 2

[14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Proceedings of the 13th European Conference on Computer Vision (ECCV), Part V*, pages 740–755, Zürich, Switzerland, 2014. Springer. 2, 3, 5

[15] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 1, 8

[16] Mengmeng Ma, Jianjie Xu, Yijie Jiang, Zhibo Wang, and Hanwang Lu. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, pages 4366–4374, 2022. 1

[17] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 3

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 2, 5, 6

[19] Ali Sharif Razavian, Hossein Azizpour, Jason Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 806–813, 2014. 2

[20] 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. 5, 6

[21] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. *arXiv preprint arXiv:2312.01597*, 2023. 2

[22] 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. *arXiv preprint arXiv:2111.07783*, 2021. 1, 2, 8

[23] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014. 2, 3, 5

[24] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language–image pre-training (siglip), 2023. 1, 2, 5, 6

[25] Guanqi Zhan, Yuanpei Liu, Kai Han, Weidi Xie, and Andrew Zisserman. Elip: Enhanced visual-language foundation models for image retrieval. In *Proceedings of the 22nd International Conference on Content-Based Multimedia Indexing (CBMI 2025)*. IEEE, 2025. 2, 3

[26] 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 *CVPR*, 2022. 3

A. Additional Experimental Details

A.1. Hyperparameter Sensitivity

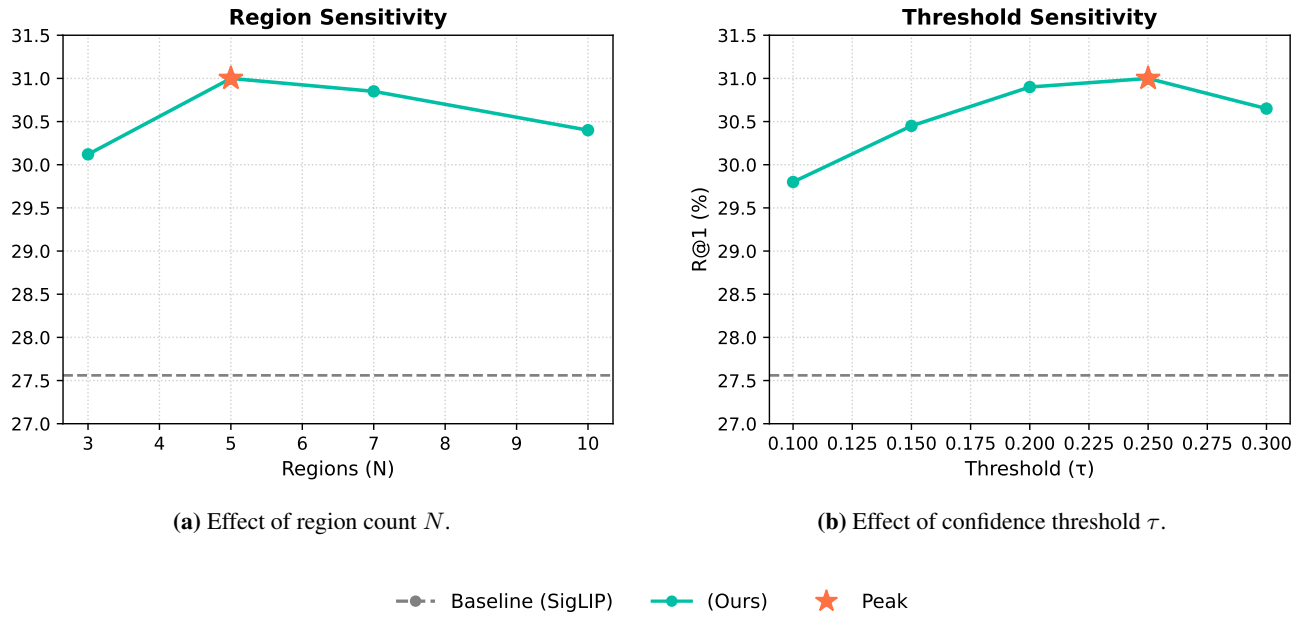
We analyze the robustness of LARE with respect to its two primary inference-time hyperparameters: the number of selected regions N and the confidence threshold τ. These parameters control the balance between computational cost and retrieval refinement. Increasing N allows the model to examine a broader set of candidate regions and improves the likelihood of recovering small or visually subtle objects that may be underrepresented in the global embedding. The threshold τ determines when regional refinement is activated, ensuring that additional computation is performed only when the global similarity signal is uncertain.

Overall, LARE remains stable across a wide range of settings and consistently improves retrieval performance over the baseline backbone. Performance increases as the number of regions grows, indicating that additional regional evidence helps resolve ambiguous queries. Beyond a moderate number of regions, gains saturate, suggesting that most relevant visual evidence has already been captured. Similarly, the method remains robust across different confidence thresholds. Based on this analysis, we use $N = 5$ and $\tau = 0.25$ throughout the paper, as this configuration provides a strong balance between retrieval accuracy and computational efficiency.

A.2. Implementation Notes

We follow the preprocessing and encoder configurations of the backbone models and use the OpenCLIP implementations of CLIP and related ViT-based encoders. All encoders remain frozen, and LARE operates entirely at inference time without modifying model parameters or requiring additional training.

For each image, we extract the self-attention tensor from an intermediate transformer layer and compute the patch-to-patch attention maps (excluding the class token). For each head, we sum each column to measure how much attention a patch receives, reshape to a spatial grid, min–max normalize, and average the top-k heads selected by spatial variance to obtain a mean attention map. We then form the inverse-attention map to identify regions that receive relatively low attention. Candidate regions are generated using a sliding window, merged using non-maximum suppression, and limited to at most N regions. Each selected region is cropped from the original image, resized to the backbone's native input resolution, and encoded using the same frozen vision encoder to obtain regional embeddings. During retrieval, LARE applies confidence-gated fusion: regional similarity is incorporated only when it provides stronger evidence than the global similarity score. This mechanism improves retrieval in dense scenes while preserving the original backbone behavior on standard benchmarks.



(a) Effect of region count N.

(b) Effect of confidence threshold τ.

Figure 5. Sensitivity of LARE to inference hyperparameters. Increasing the number of regions improves retrieval performance until saturation around $N = 5$. The method remains stable across thresholds and consistently outperforms the baseline.

B. Model Card

We provide a brief model card for LARE.

- **Model Architecture**: LARE is a training-free augmentation pipeline that operates on frozen pretrained vision-language models. The pipeline contains three main components: (1) a vision transformer encoder for extracting global image embeddings and spatial attention maps, (2) a text transformer encoder for extracting text embeddings, and (3) an inverse-attention module that detects low-attention regions, re-encodes them independently, and adaptively fuses regional and global features. The vision and text encoders are frozen pretrained models, instantiated as CLIP ViT-L/14, SigLIP SoViT-400M/14, or SigLIP 2 SoViT-400M/16, accessed via OpenCLIP.
- **Inputs**: The vision encoder takes an image as input, preprocessed to match the backbone's native resolution: $224 \times 224 \times 3$ for CLIP ViT-L/14, and $384 \times 384 \times 3$ for SigLIP and SigLIP 2 models. The text encoder takes a tokenized text string, cropped to the first 64 tokens as input.
- **Outputs**: The vision and text encoders output a d-dimensional feature vector, where d is 768 for CLIP ViT-L/14 and 1152 for SigLIP and SigLIP 2 SoViT-400M models. The pipeline outputs a fused similarity score between the text query and image.
- **Intended Use**: The method is designed for zero-shot image–text retrieval research purposes. The pipeline can be used for text-to-image and image-to-text retrieval by comparing feature vectors. The method is particularly effective for challenging retrieval scenarios where queries target fine-grained details, small objects, or background elements that may be under-emphasized by global embeddings.
- **Training Data**: LARE requires no training or fine-tuning. All vision and text encoders are frozen pretrained models (e.g., CLIP and SigLIP). The inverse-attention module operates entirely at inference time and requires no additional training data.
- **Evaluation Data**: Zero-shot retrieval is performed on MS-COCO, Flickr30k, and a curated dense-scene dataset (DenseSet) to demonstrate performance across different retrieval difficulty levels.
- **Hardware & Software**: The method is implemented in Python using PyTorch and OpenCLIP and evaluated on NVIDIA Quadro RTX 8000 GPUs (48GB).

C. Pseudocode

Algorithm 1 LARE: Low-Attention Region Encoding for Retrieval

Require: Image I, text query q, frozen vision encoder f_v, text encoder f_t, layer ℓ, top heads k, max regions N, confidence threshold τ
Ensure: Retrieval score S

Stage 1: Low-Attention Region Detection

$\{\mathbf{A}^{(h)}\}_{h=1}^{H} \leftarrow f_v(I, \ell)$ ▷ Extract attention maps at layer ℓ
for each head $h = 1, \ldots, H$ **do**
  $\mathbf{a}_i^{(h)} \leftarrow \sum_j \mathbf{A}_{j,i}^{(h)}$ for all patches i ▷ Received attention
  $\mathbf{a}^{(h)} \leftarrow \textsc{MinMaxNorm}(\mathbf{a}^{(h)})$
end for
$\mathcal{H}_k \leftarrow$ top-k heads by $\mathrm{Var}(\mathbf{a}^{(h)})$
$\bar{\mathbf{A}} \leftarrow \frac{1}{k} \sum_{h \in \mathcal{H}_k} \mathbf{a}^{(h)}$
$\mathbf{M} \leftarrow \mathbf{1} - \bar{\mathbf{A}}$ ▷ Inverse attention map
$\mathcal{W} \leftarrow \textsc{SlidingWindow}(\mathbf{M})$ ▷ Candidate windows
$\mathcal{R} \leftarrow \mathrm{NMS}(\mathcal{W})$
$\mathcal{R} \leftarrow \textsc{TopN}(\mathcal{R}, N)$ ▷ Keep top-N regions

Stage 2: Regional Encoding

for each region $r_j \in \mathcal{R}$ **do**
  $\mathbf{z}_j \leftarrow f_v(\textsc{CropAndResize}(I, r_j))$
end for

Stage 3: Confidence-Gated Scoring

$\mathbf{z}_g \leftarrow f_v(I)$; $\mathbf{z}_t \leftarrow f_t(q)$
$s_g \leftarrow \mathrm{sim}(\mathbf{z}_t, \mathbf{z}_g)$ ▷ Global similarity
$s_r \leftarrow \max_j \mathrm{sim}(\mathbf{z}_t, \mathbf{z}_j)$ ▷ Best regional match
if $s_g < \tau$ **and** $s_r > s_g$ **then**
  $\alpha \leftarrow \min\big(2(s_r - s_g),\, 0.5\big)$
  $S \leftarrow (1 - \alpha)\, s_g + \alpha\, s_r$
else
  $S \leftarrow s_g$
end if
return S