LoCaR: Localized Canvas Retrieval with Frozen Vision-Language Models

Anonymous authors
Paper under double-blind review

Abstract

Whole-image embeddings can under represent small objects and local attributes that distinguish relevant images for localized queries. We present LoCaR, a training-free framework for localized text-to-image retrieval with frozen vision–language models. LoCaR constructs and utilizes independent proposals (SAM2) focusing on the object so as to get a more localized representation that independently encodes each region using the frozen visual encoder. We further study paired-support adaptation as an optional residual that uses matched text–region exemplars without updating the underlying encoders. We also introduce LocFG, a Visual Genome-derived benchmark designed to evaluate localized retrieval for smaller and more challenging referents. Across several vision language backbones, our results show that improving the region representation itself is the main source of gain.

1 Introduction

Image search queries often describe an object or attribute within a larger scene, such as "a woman carrying a yellow handbag." Retrieving the relevant image requires preserving local evidence that may occupy only a small part of the photograph. Vision–language models such as CLIP and SigLIP2 provide efficient retrieval through aligned image and text embeddings, but a single whole-image representation can underrepresent these details (Radford et al., 2021; Tschannen et al., 2025). We study *localized image retrieval*, where a referring expression is used to rank images across a gallery. This differs from referring-expression comprehension, which localizes an object within an already supplied image.

Region representations offer a way to retain local evidence until a query is available. TextRegion constructs text-aligned region embeddings from segmentation masks and frozen image–text models, providing a strong foundation for localized matching (Xiao et al., 2025). Its region vectors aggregate features extracted from the surrounding image. An alternative is to construct a region-specific visual input and encode it independently. Cropping and background suppression have established precedents in ReCLIP and fine-grained visual prompting (Subramanian et al., 2022; Yang et al., 2023). We investigate how this alternative representation affects gallery retrieval under matched backbones and candidate proposals.

We introduce LoCaR, a frozen region-canvas retrieval framework. Each SAM2 proposal (Ravi et al., 2025) defines a padded crop in which foreground pixels are preserved and the background is blurred. The frozen vision–language model independently encodes each canvas using its standard image readout. Gallery representations are cached offline, and the maximum query–region similarity supplies an image-level retrieval score. A whole-image similarity channel provides complementary scene context. Within each backbone, comparisons with TextRegion (Xiao et al., 2025) preserve proposal identities, query embeddings, and global embeddings, measuring the combined effect of cropping, background treatment, and independent encoding.

We separately investigate paired-support adaptation through an optional residual cache. A query and candidate region are compared through their affinity profiles over matched support expressions and visual anchors. Agreement over corresponding support examples contributes additional matching evidence without updating the encoders. Applying this residual to both TextRegion and canvas representations separates its incremental benefit from the representation change.

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.

We show experiments on the dataset RefCOCOg (Mao et al., 2016). Our method LOCAR improves unadapted region R@1 on multiple Recall@k on several frozen vision backbones which are CLIP-B/16, SigLIP2-B/16 and Qwen3-VL-Embedding-2B. Our method performs significantly better with 2.40-6.05pp for CLIP, 3.13-5.94 pp for SigLIP2, and 9.13-11.69 pp for Qwen wrt zero shot backbones and beats the SOTA zero-shot retriever TextRegion on all the backbones. We show in the results that the strongest improvement comes from the local region retrieval where the attention only adheres to each region by constructing independent region canvases which enables the attention to be localized. We introduce LOCFG a Visual Genome-derived benchmark designed for smaller referents (Krishna et al., 2017). We further evaluate LOCAR on LOCFG with experiments using both transferred and in-domain support. We also adapt our method to recent backbone such as Qwen3-VL-Embedding-2B to show that LOCAR can extend to specific scenarios.

Our contributions are:

1. We introduce LOCAR, a training-free localized retrieval framework that converts segmentation proposals into independently encoded region canvases. The resulting embeddings are cached offline and can be combined with whole image similarity at retrieval time while keeping the vision–language model fully frozen.
2. We show through controlled comparisons and ablations that **region representation is the main source of improvement**. In particular, canvas construction and the text-aligned image readout have a much larger effect than paired-support adaptation which provides a smaller and setting-dependent correction.
3. We introduce LOCFG, a small-referent retrieval benchmark, and use RefCOCOg and LOCFG to study how region representation, support domain and global–local fusion affect localized retrieval across CLIP, SigLIP2 and Qwen backbones.

2 RELATED WORK

Frozen region representations and visual prompting. TextRegion (Xiao et al., 2025) constructs text-aligned region tokens by aggregating frozen visual features under SAM2 masks, providing our principal representation baseline. An alternative is to modify the visual input before encoding. ReCLIP (Subramanian et al., 2022) scores object proposals using cropped and background-blurred views with frozen CLIP. Fine-Grained Visual Prompting (FGVP) (Yang et al., 2023) uses SAM masks to preserve a target's foreground while blurring the surrounding background on the full image. These methods establish cropping and foreground-preserving background suppression as useful visual prompts. LOCAR combines padded proposal crops with mask-guided background blur and independently encodes the resulting canvases for gallery retrieval. Our matched comparison with TextRegion measures the combined representation change while holding the backbone and candidate proposals fixed.

Support-based adaptation. Tip-Adapter (Zhang et al., 2022) augments zero-shot CLIP (Radford et al., 2021) classification scores with a cache of support-image features and one-hot class labels. CLIP-Adapter (Gao et al., 2024) learns lightweight feature adapters, CoOp and CoCoOp (Zhou et al., 2022a;b) optimize text prompts, and APE (Zhu et al., 2023) refines pretrained features for few-shot classification. Our optional paired-support residual uses matched expression–region anchors rather than class-label values. Text and visual affinity profiles retain corresponding support-pair indices, allowing their agreement to contribute to retrieval scores. This construction accommodates free-form queries without specifying a fixed classification vocabulary. The encoders remain frozen, and setting the residual weight to zero recovers the corresponding base region scorer. We evaluate the residual on both TextRegion and canvas representations to distinguish support adaptation from representation changes.

Localized retrieval benchmarks. Referring-expression datasets such as RefCOCO and RefCOCOg (Mao et al., 2016; Yu et al., 2016) associate descriptions with objects inside individual images. REIR (Hao et al., 2025) extends this setting to retrieval across images with instance localization. SORCE (Liu et al., 2025) evaluates image retrieval from small-object descriptions and studies both frozen regional prompting and fine-tuned embeddings. ILIAS (Kordopatis-Zilos et al., 2025) evaluates instance-level retrieval against large distractor galleries. Our evaluation ranks gallery images

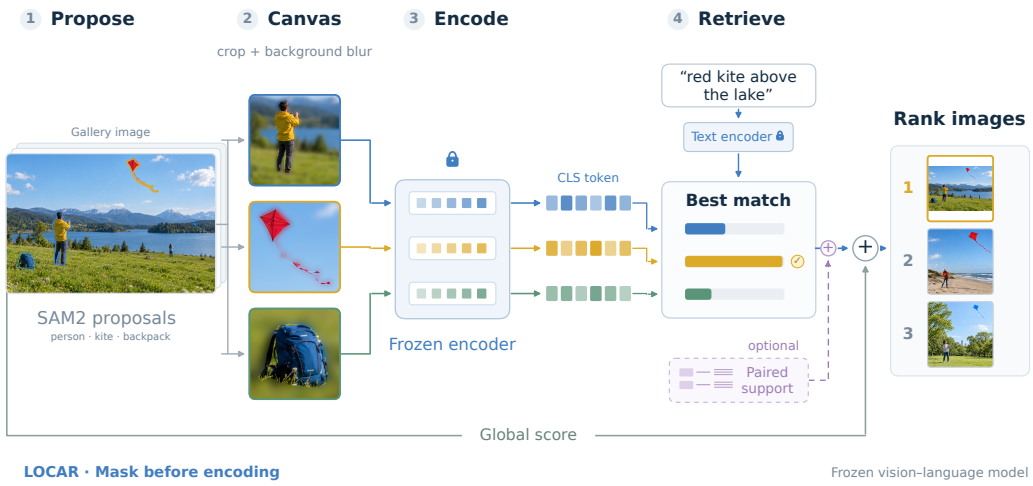

Figure 1: **Region-canvas retrieval.** (1) SAM2 generates candidate regions. (2) Each proposal defines a foreground-preserving crop with blurred background. (3) A shared frozen encoder independently embeds each canvas, and the embeddings are cached. (4) Query–region matching, optional paired-support evidence, and whole-image similarity produce the image ranking.

from localized descriptions and measures image-retrieval recall, without requiring a predicted box to satisfy a localization threshold. We additionally construct LocFG from Visual Genome (Krishna et al., 2017) to examine retrieval for smaller referents, with image-disjoint fit and test partitions and transferred versus in-domain support.

3 Region-Canvas Retrieval

Figure 1 summarizes our pipeline: generate proposals, construct region canvases, encode them with a frozen vision–language model, and retrieve images using cached representations. Each image receives a maximum query–region similarity score, optionally augmented by paired-support evidence and combined with whole-image similarity. The vision–language encoders remain frozen throughout.

3.1 Task and Proposal Cache

Let $\mathcal{G}$ denote the image gallery and T a query describing a referent. Let f_t and f_v be the text and image embedding functions of a fixed vision–language model. We define $\mathrm{nrm}(x) = x/\|x\|_2$ and obtain the normalized query embedding as $t = \mathrm{nrm}(f_t(T))$.

SAM2 (Ravi et al., 2025) generates class-agnostic masks and bounding boxes for each gallery image. Let N_I denote the number of cached SAM2 proposals available for image I, before applying the scoring cap described below. We reuse the same proposal identities and ordering for TextRegion and canvas representations. Proposals may overlap or represent an object and its parts; they need not partition the image. Gallery proposals and embeddings are constructed without query annotations or ground-truth referent boxes.

3.2 Constructing and Encoding Region Canvases

Stages 2–3 of Figure 1 construct and encode a separate visual input for each proposal. We map the proposal box from the cached mask coordinate frame to the source image, expand it by 10% of its width and height on each side, enforce a minimum side length of eight pixels, and clip it to the image boundary. Let J_i denote the resulting RGB crop and $\widetilde{m}_i$ its aligned binary mask. Cached masks are stored within their tight proposal boxes at a maximum long-side resolution of 64 pixels. We reconstruct each mask by nearest-neighbor resizing into its proposal footprint and assign zero mask values to the added crop margin.

The canvas preserves foreground pixels and blurs the surrounding background:

$$C_i = \widetilde{m}_i \odot J_i + (1 - \widetilde{m}_i) \odot \mathcal{B}_{55}(J_i), \tag{1}$$

where $\odot$ broadcasts the mask over RGB channels and $\mathcal{B}_{55}$ denotes Gaussian blur with a 55×55 kernel. We use OpenCV's kernel-derived default standard deviation and apply blurring at the crop's original resolution, before resizing the composite to 224×224 for CLIP-B/16 and SigLIP2-B/16. The blurred background retains coarse context while suppressing detail outside the foreground mask.

Let P denote resizing followed by the backbone's image normalization. The region embedding is

$$r_i = \mathrm{nrm}(f_v(P(C_i))) \,. \tag{2}$$

We use the model's standard text-aligned image readout: the projected class token for CLIP-B/16 and SigLIP2-B/16. Canvases are encoded independently, with batching permitted. Thus, visual attention operates on each constructed canvas before a region vector is produced.

For Qwen3-VL-Embedding-2B, we reuse the cached SAM2 proposal boxes and independently encode each plain RGB crop through the model's image-embedding interface, without background suppression. This avoids requiring a patch-to-text projection for mask pooling. The normalized crop embeddings use the same retrieval and paired-support scoring as the canvas embeddings. For in-domain LOCFG evaluation, the query instruction "Find the photo containing the part of the scene that this phrase describes" is selected on fit data before test evaluation (Appendix A.4).

We cache proposal embeddings in their original order and retain the first $R_I = \min(N_I, 32)$ candidates for scoring:

$$\mathcal{R}(I) = \{r_i\}_{i=1}^{R_I}. \tag{3}$$

Here, N_I counts all available cached proposals for image I, whereas R_I counts those used by the retrieval scorer. The cache may therefore contain more embeddings than are scored. Gallery candidates are constructed independently of support matching. Independent encoding increases offline visual computation relative to pooling shared image features; the cached embeddings are reused for all queries.

3.3 REGION-BASED IMAGE SCORING

The retrieval stage in Figure 1 compares the query with every retained region and assigns each image its largest query–region cosine similarity:

$$s_0(T, I) = \max_{r \in \mathcal{R}(I)} \langle t, r \rangle. \tag{4}$$

We rank gallery images by this score. Regions supply evidence for image retrieval; the relevant image is not provided in advance. The base scorer requires no support examples. Its effectiveness depends on the retained proposals: a missing or poorly framed referent may not be represented adequately by any canvas.

3.4 OPTIONAL PAIRED-SUPPORT RESIDUAL

We optionally add a LOCAR+ residual inspired by support-cache adaptation (Zhang et al., 2022). For each labeled support expression, we match its ground-truth box to the retained proposal with the highest box IoU after aligning their coordinate frames. We discard matches below the threshold τ_{IoU}. The normalized expression embedding and matched visual embedding form a support pair (u_j, v_j) in $\mathcal{B} = \{(u_j, v_j)\}_{j=1}^{M}$. We retain at most $4{,}096$ pairs, subsampling with a fixed seed when necessary. Visual anchors use the same backbone and region representation as the gallery candidates.

For $\beta > 0$, affinities to the text and visual anchors are

$$\begin{aligned} a_t(t)_j &= \exp[\beta(\langle t, u_j \rangle - 1)] \,, \\ a_r(r)_j &= \exp[\beta(\langle r, v_j \rangle - 1)] \,. \end{aligned} \tag{5}$$

We sparsify and normalize these affinities:

$$\begin{aligned} \phi_t(t) &= \mathrm{nrm}(\mathrm{TopK}(a_t(t))) \,, \\ \phi_r(r) &= \mathrm{nrm}(\mathrm{TopK}(a_r(r))) \,. \end{aligned} \tag{6}$$

The TopK operator retains the K largest entries at their original indices and sets all others to zero. Both profiles therefore use the same M support-pair coordinates, and their inner product measures agreement over matched exemplars.

The adapted score is

$$s_\gamma(T, I) = s_0(T, I) + \gamma \max_{r \in \mathcal{R}(I)} \langle \phi_t(t), \phi_r(r) \rangle, \qquad \gamma \geq 0. \tag{7}$$

The direct-similarity and support-agreement terms are maximized independently and may select different proposals. Setting $\gamma = 0$ exactly recovers the base region scorer. Applying the same residual to TextRegion embeddings gives the TextRegion + configuration evaluated in Sec. 4.5.

3.5 Global fusion and calibration

Let $g(I)$ be the normalized whole-image embedding and define $s_{\mathrm{glob}}(T, I) = \langle t, g(I) \rangle$. Within each backbone, representation comparisons reuse the same cached text and whole-image embeddings; canvas construction changes only the region vectors. For a fixed query, $z_T[s](I)$ standardizes channel s using its mean and standard deviation across the gallery. The fused score is

$$s_{\mathrm{fuse}}(T, I) = z_T[s_{\mathrm{glob}}](I) + \alpha \, z_T[s_\gamma](I). \tag{8}$$

Support images are disjoint from the evaluation gallery. For calibration, we divide support images into fit and held-out halves and use fit anchors to score held-out queries and images. Affinity sharpness β is derived from support similarity statistics. We select K and γ using held-out retrieval performance, then select α conditional on that residual configuration. The base and adapted rows within each comparison share this α, measuring the residual's effect at a fixed fusion weight. All selected settings are fixed before test evaluation.

Transferred evaluation retains the RefCOCOg support bank and calibration. In-domain evaluation constructs the bank and selects hyperparameters using the LocFG fit partition. Consequently, unadapted fused scores can differ across settings through α. Unadapted region scores remain identical when query instructions and visual representations are fixed; Qwen's settings additionally differ in their query instructions..

4 Evaluation

4.1 Experimental setup

We evaluate image retrieval on RefCOCOg (Mao et al., 2016) ($9{,}602$ queries, $2{,}600$ gallery images) and introduce LocFG, a Visual Genome-derived benchmark (Krishna et al., 2017) for retrieval from descriptions of small referents. RefCOCOg provides an established evaluation setting, but our audit of the public mirror revealed split-resolution and validation test overlap issues can contaminate support-set construction when provenance is inferred only from nominal split labels (Appendix B). LocFG complements this evaluation in two ways: it tests retrieval when the described object occupies little of the image, and it provides independently constructed image-disjoint fit/test partitions with overlap filtering against RefCOCOg. It contains $12{,}033$ test queries over $2{,}600$ gallery images; the median referent bounding box occupies 3.0% of image area, compared with 14.5% on RefCOCOg. This setting tests whether region canvases preserve evidence about small targets and whether paired-support gains generalize beyond RefCOCOg, using either transferred RefCOCOg support or in-domain LocFG support. Construction and filtering details appear in Appendix C.

We use frozen CLIP-B/16 (Radford et al., 2021), SigLIP2-B/16 (Tschannen et al., 2025), and Qwen3-VL-Embedding-2B (Li et al., 2026) encoders as backbones for evaluating LocAR. Matched-representation comparisons reuse the same SAM2 (Ravi et al., 2025) proposals, and hyperparameters are selected on held-out support images before test evaluation. Details of benchmark construction, support filtering, proposal settings, and calibration are provided in Appendix A.1; the support-set contamination audit is reported in Appendix B.

4.2 Representations, baselines, and metrics

We use *TextRegion* (Xiao et al., 2025) as our strongest directly comparable training-free region baseline. Its published CLIP configuration surpasses the reported zero-shot RefCOCOg referring-

Table 1: **RefCOCOg retrieval** (%). The right-hand columns report standalone global recall for zero-shot rows and global–region fusion otherwise. "+" adds paired support. Best and second-best values are marked within each backbone and metric. Standalone and matched fusion global channels are compared in Appendix A.2.

		Region retrieval			Global / fused		
Configuration	**Model**	R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot (Global)	CLIP-B/16	–	–	–	15.15	32.83	42.89
TextRegion		9.66	26.25	37.94	17.44	37.40	48.33
TextRegion +		10.42	**27.88**	**39.71**	**17.52**	**37.61**	48.81
LoCaR		**12.04**	27.47	37.53	17.38	37.49	**48.91**
LoCaR +		**12.71**	**28.70**	**39.30**	**17.57**	**37.73**	**49.04**
Zero-shot (Global)	SigLIP2-B/16	–	–	–	22.80	43.58	53.76
TextRegion		9.57	24.68	35.53	24.69	46.66	57.63
TextRegion +		10.10	26.40	37.70	24.88	47.16	58.05
LoCaR		**19.26**	**39.19**	**49.51**	**25.99**	**48.48**	**59.54**
LoCaR +		**19.95**	**40.63**	**51.42**	**25.93**	**48.45**	**59.70**
Zero-shot (Global)	Qwen3-VL-2B	–	–	–	20.17	40.56	51.92
LoCaR		**19.90**	**37.98**	**47.80**	**29.15**	**52.60**	**63.11**
LoCaR +		**20.67**	**39.39**	**49.50**	**29.30**	**52.85**	**63.61**

expression comprehension results of RedCircle (Shtedritski et al., 2023), FGVP (Yang et al., 2023), and CPT (Yao et al., 2024). TextRegion constructs text-aligned region embeddings by pooling patch value features with proposal-mask weights through the frozen backbone's readout. Here, we evaluate these representations for gallery-level image retrieval. Comparisons with LoCaR share the backbone, proposals, query embeddings, and whole-image embeddings, while changing how each proposal is represented.

LoCaR independently encodes each proposal as a canvas that preserves the foreground and blurs the surrounding background (Appendix A.1). For Qwen, we use plain proposal crops because its embedding interface does not expose the patch-to-text projection required by our TextRegion implementation. "+" denotes paired expression region support: *TextRegion* + applies the residual to mask-pooled features, whereas LoCaR + applies it to independently encoded regions.

Global zero-shot ranks gallery images using whole-image cosine similarity to the query. *Region* uses the local image score $s_\gamma(T, I)$ from Eq. 7: the maximum query-region cosine similarity, optionally augmented by the paired-support residual. Without support, $\gamma = 0$ and this reduces to $s_0(T, I)$ in Eq. 4. *Fused* combines the global and local channels after per-query z-score standardization: each channel's mean across gallery images is subtracted and the result is divided by its standard deviation (Eq. 8). Each matched pair of results without and with support uses the same fusion weight within its representation and support setting.

We report R@1, R@5, and R@10 (%), the percentages of queries whose source image appears among the top 1, 5, and 10 retrieved gallery images, respectively. Gains are expressed in percentage points (pp), unless stated otherwise. Evaluation and calibration details, including the distinction between standalone and matched global channels, appear in Appendices A.1 and A.2.

4.3 Main results on RefCOCOg

Overall retrieval performance. Table 1 shows that LoCaR + achieves the highest fused recall among the evaluated configurations with Qwen: $29.30/52.85/63.61\%$ at R@1/R@5/R@10. Relative to the standalone global zero-shot rows, fused LoCaR + improves these metrics by $2.42/4.90/6.15$ pp for CLIP, $3.13/4.87/5.94$ pp for SigLIP2, and $9.13/12.29/11.69$ pp for Qwen. These comparisons establish gains over standalone retrieval, but do not isolate fusion when the global configurations differ. Against the matched global channels used in fusion, the R@1 gains are 2.71, 3.13, and 2.82 pp, respectively (Appendix A.2).

Table 2: **LocFG retrieval** (%). "+" adds paired support. Tr and ID denote transferred and in-domain settings, including their calibration choices. Identical baseline values are shown once across both columns. Separate baseline values reflect different fusion calibration or, for Qwen, different query instructions. Global zero-shot reports standalone retrieval; other right-hand entries report global–region fusion. Best and second-best values are marked within each backbone and metric across both settings. Qwen settings are detailed in Appendix A.4.

	Region retrieval						**Global / fused**					
	R@1		R@5		R@10		R@1		R@5		R@10	
Configuration	Tr	ID	Tr	ID	Tr	ID	Tr	ID	Tr	ID	Tr	ID
(a) CLIP-B/16												
Global zero-shot	–		–		–		7.01		15.80		20.76	
TextRegion	5.68		13.71		18.73		8.23		18.39		24.76	
TextRegion +	5.76	5.80	13.79	13.84	18.79	18.86	8.30	8.30	18.42	18.42	24.79	24.79
LoCaR	6.07		14.42		19.47		**8.44**		**18.81**		**24.77**	
LoCaR +	**6.24**	**6.26**	**14.85**	**14.96**	**20.01**	**20.15**	8.41	**8.76**	18.80	**19.31**	24.68	**25.85**
(b) SigLIP2-B/16												
Global zero-shot	–		–		–		10.20		21.09		27.32	
TextRegion	7.63		17.50		23.50		12.00		25.00		32.00	
TextRegion +	7.71	7.75	17.58	17.53	23.62	23.79	12.05	12.12	25.05	25.05	32.05	32.05
LoCaR	9.51		20.84		27.02		**12.04**		**25.45**		**33.11**	
LoCaR +	**9.71**	**9.57**	**21.33**	**21.41**	**27.87**	**27.61**	12.02	**12.53**	25.19	**26.56**	33.05	**33.87**
(c) Qwen3-VL-2B												
Global zero-shot	–		–		–		11.78		23.94		31.16	
LoCaR	**12.27**		**26.43**		**33.91**		**14.32**		**29.55**		**38.34**	
LoCaR +	11.08	**12.46**	24.05	**26.71**	31.53	**34.26**	13.73	**14.40**	28.31	**29.60**	37.09	**38.38**

Region representation and paired support. Without paired support, LoCaR improves region R@1 over TextRegion by 2.38 pp for CLIP and 9.69 pp for SigLIP2. Adding support increases these total gains to 3.05 and 10.38 pp; the residual itself contributes 0.67 and 0.69 pp. Thus, the larger improvement comes from the canvas representation. At R@10, LoCaR + exceeds unadapted TextRegion by 1.36 pp for CLIP and 15.89 pp for SigLIP2. The advantage is not uniform: CLIP TextRegion + retains a 0.41 pp lead over LoCaR + at region R@10. For Qwen's crop representation, paired support adds $0.77/1.41/1.70$ pp at R@1/R@5/R@10.

Complementarity with global retrieval. Fused LoCaR + also outperforms fused TextRegion at all three ranks: by $0.13/0.33/0.71$ pp for CLIP and $1.24/1.79/2.07$ pp for SigLIP2. The stronger region representation therefore contributes beyond both global retrieval and the TextRegion baseline, particularly with SigLIP2. However, region improvements are attenuated by fusion. For SigLIP2, adding support improves region recall at every rank but slightly reduces fused R@1/R@5 by $0.06/0.03$ pp, while increasing fused R@10 by 0.16 pp. Support improves all three fused metrics for all backbones. Its benefit therefore depends on how the local scores complement the global channel, rather than on region recall alone.

Why do the gains differ across backbones? CLIP benefits from LoCaR, but its smaller region gains leave less improvement to carry into fusion. Canvas construction changes the visual input before encoding, restricting attention to the crop and suppressing background detail. TextRegion instead pools region features that already contain context from the source image. The results are consistent with canvas encoding providing more discriminative local evidence, especially for SigLIP2. However, this comparison jointly changes cropping, background treatment, and feature aggregation; it does not isolate attention locality or establish pretraining objectives as the cause of the backbone differences.

4.4 LocFG: transferred and in-domain support

We evaluate generalization on LocFG, our small-referent retrieval benchmark introduced in Sec. 4.1. All results in Table 2 use the same LocFG test queries and gallery. The transferred setting (Tr) reuses the RefCOCOg support bank and calibration, whereas the in-domain setting (ID) uses support and calibration from the image-disjoint LocFG fit partition. This comparison tests whether support from another dataset remains useful and whether adapting to the target domain provides additional benefit.

Rows without "+" are unadapted configurations: they use no paired-support residual, corresponding to $\gamma = 0$. Rows with "+" add the paired expression–region residual. For a fixed representation and query instruction, unadapted region scores are independent of the support bank. Fused scores can nevertheless differ across settings because the calibrated fusion weight changes.

The three backbone panels include standalone global retrieval and LoCaR with and without paired support. CLIP and SigLIP2 additionally include TextRegion and TextRegion +. Qwen uses plain proposal crops, and its transferred and in-domain runs use different query instructions. Consequently, the Qwen comparison reflects changes in both instruction and support configuration (Appendix A.4).

Overall retrieval performance. The strongest reported LocFG configuration is in-domain Qwen LoCaR +, reaching region recall of $12.46/26.71/34.26\%$ and fused recall of $14.40/29.60/38.38\%$ at R@1/R@5/R@10. Relative to the standalone global rows, in-domain fused LoCaR + improves these metrics by $1.75/3.51/5.09$ pp for CLIP, $2.33/5.47/6.55$ pp for SigLIP2, and $2.62/5.66/7.22$ pp for Qwen. Thus, the complete configurations improve retrieval beyond standalone global embeddings on all three backbones. For Qwen, this standalone comparison also includes differences in the query configuration; matched global comparisons are reported in Appendix A.4.

Representation gains and transferred support. Without paired support, LoCaR improves region R@1/R@10 over TextRegion by $0.39/0.74$ pp for CLIP and $1.88/3.52$ pp for SigLIP2. Transferred LoCaR + increases these total gains to $0.56/1.28$ and $2.08/4.37$ pp, respectively. The transferred residual therefore improves region recall at all three ranks for these two backbones. However, its fused R@1/R@10 effects are slightly negative: $-0.03/-0.09$ pp for CLIP and $-0.02/-0.06$ pp for SigLIP2. For Qwen, transferred support raises region R@1 by 0.12 pp but lowers R@5 by 0.08 pp. Transferred support can improve local matching, but its benefit is neither uniform across metrics nor guaranteed to improve fusion.

What does in-domain support add? In-domain LoCaR + exceeds TextRegion at region R@1/R@10 by $0.58/1.42$ pp for CLIP and $1.94/4.11$ pp for SigLIP2. Relative to LoCaR without support, the region R@1 increments are modest: 0.19 pp for CLIP, 0.06 pp for SigLIP2, and 0.19 pp for Qwen. In-domain LoCaR + achieves higher fused recall than its transferred counterpart at all three ranks for every backbone. This does not establish that the support bank alone is responsible: fusion calibration changes, and Qwen additionally changes its query instruction. Nor is in-domain support uniformly stronger for region retrieval: SigLIP2's transferred LoCaR + retains higher R@1 and R@10 than its in-domain counterpart.

Interpreting the smaller gains. The smaller CLIP and SigLIP2 region improvements than on RefCOCOg show that the benefit of canvas encoding depends on the evaluation setting. LocFG emphasizes smaller referents: the median annotated box occupies 3.0% of image area, compared with 14.5% on RefCOCOg. At our operating point, only 51.1% of LocFG queries have a retained proposal with box IoU ≥ 0.5 against the referent. Limited availability of well-localized candidates may therefore constrain region encoding, although this coverage is not an upper bound on image-retrieval recall. Together, the results motivate better small-region proposals and representations; they do not isolate object size, proposal quality, or pretraining as the sole cause of the cross-dataset differences.

4.5 Ablation studies

In this section we examine various canvas masking, pooling startegies, text instructions template and cache capacity usage. Details on calibration settings, cache sizes and Qwen instruction selection are in Appendix A.1, A.5, A.4.

Background masking LoCaR constructs region embeddings by independently encoding each SAM2 proposal as a canvas. The foreground region is preserved but the surrounding background is suppressed and the composite image is passed through the frozen image encoder. A key design choice

is how aggressively to suppress the background (a) soft suppression via Gaussian blur keeps the surround visually plausible and in-distribution for the encoder (b) while hard suppression via white fill removes all contextual information at the cost of introducing an out-of-distribution background. We ablate these two treatments on SigLIP2-B/16 with the RefCOCOg test split as shown in 3. White fill reduces region R@1 by 2.72 pp versus blur but improves fused R@1 by 0.30 pp, likely because stronger background removal makes region errors more complementary to the global embedding. Blur preserves more context and gives stronger standalone region representations, while white fill provides slightly better fusion. Since blur still outperforms TextRegion by +9.69 pp region R@1 and +1.30 pp fused R@1, we use blur as the default; the white-fill result mainly highlights the tradeoff between region quality and global-region complementarity.

Proposal vector pooling After encoding each canvas crop with the frozen ViT, we compare two region readouts in 4 the attention-pooled CLS representation and mean pooling over patch tokens. Mean pooling performs substantially worse: region R@1 improves by only +0.75 pp over TextRegion, compared with +9.69 pp using CLS, while fused R@1 falls below TextRegion (23.67% vs. 24.69%). A likely reason is that the patch tokens are intermediate visual features and are not individually optimized to align with text. Simply averaging them therefore does not necessarily produce a meaningful point in the joint image–text embedding space. In contrast, the CLS readout is explicitly optimized through the model's contrastive objective to represent the image in that space. Paired support adds only +0.30 pp with mean pooling, further suggesting that support cannot fully compensate for a poorly aligned base representation. Overall, this ablation shows that how the canvas is pooled is critical: independent encoding is most effective when its text-aligned CLS readout is preserved.

Table 3: **SigLIP2 background ablation** on RefCOCOg (%). Both configurations use the default attention-pooling readout, labeled `cls` in the experiment logs. The shared global channel gives $22.80/43.58/53.76$ at R@1/R@5/R@10.

		Region			**Fused**		
Background	Support	R@1	R@5	R@10	R@1	R@5	R@10
Blur	Without	19.26	39.19	49.51	25.99	48.48	59.54
	With	19.95	40.63	51.42	25.93	48.45	59.70
White	Without	16.54	33.49	43.43	26.29	49.10	60.06
	With	17.38	35.75	46.35	26.17	49.11	60.30

Table 4: **SigLIP2 mean-pooling ablation** on RefCOCOg (%). Region canvases use blurred backgrounds and mean pooling. Support denotes the paired-support residual. The global channel is identical to that in Table 3.

	Region			**Fused**		
Support	R@1	R@5	R@10	R@1	R@5	R@10
Without	10.32	26.23	36.63	23.67	45.41	56.01
With	10.62	26.60	36.93	23.72	45.46	56.08

5 CONCLUSION

We presented LOCAR, a training-free approach that independently encodes proposal-based region canvases with frozen vision–language models and combines local evidence with whole-image similarity. Experiments on RefCOCOg and our small-referent benchmark LOCFG show that improving region representations provides larger gains than the additional paired-support residual. Across the backbones LOCAR + shows lift of 2.40-9.13 pp on recall metrics on RefCOCOg. Ablations highlight the importance of background treatment and visual readout. Future work will focus on better small-referent proposals, support adaptation under domain shift, and more efficient region encoding and caching.

AI USE STATEMENT

Generative AI tools were used for language editing, LaTeX restructuring, and the initial layout of explanatory figures. All scientific claims, equations, experimental results, statistical tests, and bibliographic entries were checked by the authors against the experiment artifacts and cited sources. No AI system selected experimental outcomes or replaced human verification of the reported numbers.

REFERENCES

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International journal of computer vision*, 132(2):581–595, 2024.

Xiangzhao Hao, Kuan Zhu, Hongyu Guo, Haiyun Guo, Ning Jiang, Quan Lu, Ming Tang, and Jinqiao Wang. Referring expression instance retrieval and a strong end-to-end baseline. pages 4464–4473, 2025.

Giorgos Kordopatis-Zilos, Vladan Stojnić, Anna Manko, Pavel Suma, Nikolaos-Antonios Ypsilantis, Nikos Efthymiadis, Zakaria Laskar, Jiri Matas, Ondrej Chum, and Giorgos Tolias. Ilias: Instance-level image retrieval at scale. pages 14777–14787, 2025.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. 2026. URL https://arxiv.org/abs/2601.04720.

Chunxu Liu, Chi Xie, Xiaxu Chen, Wei Li, Feng Zhu, Rui Zhao, and Limin Wang. Sorce: Small object retrieval in complex environments. *arXiv preprint arXiv:2505.24441*, 2025.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

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.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. 2025:28085–28128, 2025.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11953–11963. IEEE, 2023.

Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. ReCLIP: A strong zero-shot baseline for referring expression comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5198–5215. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.357. URL https://aclanthology.org/2022.acl-long.357/.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Yao Xiao, Qiqian Fu, Heyi Tao, Yuqun Wu, Zhen Zhu, and Derek Hoiem. Textregion: Text-aligned region tokens from frozen image-text models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=KZLmkL62M4. J2C Certification.

Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 24993–25006, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4e9fa6e716940a7cfc60c46e6f702f52-Abstract-Conference.html.

Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *AI Open*, 5:30–38, 2024.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pages 69–85. Springer, 2016.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 16795–16804. IEEE, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022b.

Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In *IEEE/CVF International Conference on Computer Vision*, 2023.

A Extended evaluation and ablations

A.1 Implementation, support filtering, and metrics

Data and support separation. RefCOCOg test contains 9,602 expressions over 2,600 gallery images. Removing 481 validation rows shared with test leaves 6,843 support candidates. We retain up to 4,096 matched expression–region pairs and audit each bank for evaluation overlap. Hyperparameters are selected on held-out support images before test evaluation. Transferred evaluation reuses the RefCOCOg bank and calibration; in-domain LocFG evaluation uses its fit partition.

LocFG is derived from Visual Genome (Krishna et al., 2017) and contains 12,033 test queries over 2,600 images. Its fit/test partitions are image-disjoint, and a perceptual-hash blocklist filters overlap with RefCOCOg..

Referent size and proposal coverage. The median referent occupies 3.0% of image area on LocFG, compared with 14.5% on RefCOCOg. At the evaluated proposal settings, 51.1% of LocFG queries have a proposal with box IoU ≥ 0.5 against the annotated referent. This measures proposal localization coverage, not an upper bound on source-image retrieval recall. The coverage-stratified experiments do not establish coverage as the sole explanation for the small support gains.

Proposals and matched comparisons. SAM2 `hiera-large` uses 16 points per side, and scoring retains at most the first 32 proposals per image. Within CLIP and SigLIP2, representation comparisons share proposal identities and ordering, query embeddings, and whole-image embeddings. TextRegion provides a frozen mask-pooled comparator; it is not required to construct or score LocAR canvases. Qwen independently encodes proposal crops because its interface does not expose the patch-to-text projection used by our TextRegion implementation.

Canvas construction. Each proposal box is expanded by 10% on each side. The canvas preserves pixels inside the proposal mask and blurs the surrounding background with a 55×55 Gaussian kernel. The composite is resized to 224×224 and independently encoded by the frozen CLIP or SigLIP2 image encoder. White-fill controls replace the background with white. The default readout is labeled `cls` in the experiment logs; this label does not imply identical pooling architectures across backbones. The mean variant uses mean pooling. Comparing canvases with TextRegion jointly changes cropping, resizing, background treatment, and independent encoding.

Scoring and calibration. For normalized query and region embeddings t and r, the unadapted region score is

$$s_0(T, I) = \max_{r \in \mathcal{R}(I)} \langle t, r \rangle.$$

The paired-support residual compares query affinities to support expressions with region affinities to their paired visual anchors, yielding s_γ. Setting $\gamma = 0$ recovers s_0. For a normalized whole-image embedding $g(I)$, $s_{\text{glob}}(T, I) = \langle t, g(I) \rangle$, and fusion uses

$$s_{\text{fuse}}(T, I) = z_T[s_{\text{glob}}](I) + \alpha z_T[s_\gamma](I).$$

For each query and channel, z_T subtracts the gallery mean and divides by the gallery standard deviation. Unadapted region scores require no support, whereas unadapted fused scores use a calibrated α. Matched comparisons without and with support share α within each representation and setting. On LocFG, SigLIP2 canvas fusion uses $\alpha = 0.5$ under transfer and $\alpha = 1.0$ in-domain; cross-setting fused comparisons therefore include calibration changes.

A.2 Global baselines and fusion gains

Tables 1 and 2 report the main retrieval comparisons. Table 5 collects standalone global results. These differ from the matched global channels used in the RefCOCOg fusion runs, whose R@1 values are 14.86, 22.80, and 26.48 for CLIP, SigLIP2, and Qwen, respectively. Subtracting a standalone global result from a fused result does not isolate fusion unless prompts, preprocessing, and scoring are matched.

Table 5: **Standalone global zero-shot retrieval** (%).

Dataset	Backbone	R@1	R@5	R@10
RefCOCOg	CLIP-B/16	15.15	32.83	42.89
	SigLIP2-B/16	22.80	43.58	53.76
	Qwen3-VL-2B	20.17	40.56	51.92
LocFG	CLIP-B/16	7.01	15.80	20.76
	SigLIP2-B/16	10.20	21.09	27.32
	Qwen3-VL-2B	11.78	23.94	31.16

Table 6: **Qwen instruction comparison on LocFG.** Region triplets are R@1/R@5/R@10 (%); gains are percentage points. Global and fused columns report matched unadapted R@1.

Instruction	Region recall	Support gain	Global R@1	Fused R@1
Incumbent (I)	10.96/24.13/31.52	+0.23/+0.43/+0.52	10.90	13.64
Selected (S)	12.27/26.43/33.91	+0.18/+0.28/+0.36	11.70	14.32

Relative gains and backbone comparisons. On RefCOCOg, fused LoCaR + exceeds matched global R@1 by 2.71 pp for CLIP, 3.13 pp for SigLIP2, and 2.82 pp for Qwen, corresponding to relative gains of 18.2%, 13.7%, and 10.6%. The support-only region R@1 gain is $0.67/12.04 = 5.6\%$ for CLIP and $0.77/19.90 = 3.9\%$ for Qwen. Their fused R@10 increments are 0.13 and 0.50 pp, respectively. Thus, Qwen's larger fused R@10 increment does not imply a uniformly larger relative region improvement. Aggregate recalls alone establish neither complementary error patterns nor a statistical ordering across backbones.

A.3 Canvas background and pooling ablations

Tables 3 and 4 compare background treatment and readout for SigLIP2-B/16 on RefCOCOg. Blur with the default readout uses $\gamma = 0.10$, $K = 64$, and $\alpha = 0.5$; white uses $\gamma = 0.10$, $K = 32$, and $\alpha = 0.5$; blur with mean pooling uses $\gamma = 0.05$, $K = 64$, and $\alpha = 0.25$. Unadapted rows set $\gamma = 0$. Unadapted region comparisons therefore exclude support and fusion, whereas adapted and fused comparisons include these calibrated settings.

Background treatment. Without support, blur exceeds white fill in region R@1/R@5/R@10 by $2.72/5.70/6.08$ pp. With support, the differences are $2.57/4.88/5.07$ pp. However, white achieves higher fused R@1: 26.29 versus 25.99 without support and 26.17 versus 25.93 with support. The reported McNemar values ($p = 0.53$ for white and $p = 0.71$ for blur) compare predictions without and with support within each setting. They do not test white against blur or establish equivalence between backgrounds.

Readout and fusion. Under blur, replacing the default readout with mean pooling reduces unadapted region R@1 by 8.94 pp, from 19.26 to 10.32. Adapted fused R@1 is also lower, at 23.72 versus 25.93, although this comparison changes γ and α as well as the readout. Z-score standardization removes channel offsets and scales; it does not improve within-channel rankings. Aggregate recalls alone do not establish whether similar fused scores arise from complementary errors or dominance of the global channel.

A.4 Qwen crops and instruction selection on LocFG

The incumbent instruction (I) is `asym-image`, and the fit-selected instruction (S) is `asym-scene-part`. The instruction comparison fixes visual features and uses $\gamma = 0.05$, $K = 16$, and $\alpha = 1$. Instructions are selected using fit data. Table 6 reports the original instruction-ablation results, with unadapted region recalls, support increments, and matched global and fused R@1.

Selecting S improves unadapted region R@1/R@5/R@10 by $1.31/2.30/2.39$ pp, matched global R@1 by 0.80 pp, and unadapted fused R@1 by 0.68 pp. The incremental support gains decrease by

Table 7: **Qwen instruction slate on LocFG.** Unadapted region recalls (%) and paired-support gains (pp). Instructions are selected on fit data.

Instruction	β	R@1/R@5/R@10	Gain R@1/R@5/R@10
Model default	12.72	8.22/18.06/23.73	+0.14/+0.91/+1.30
`asym-object`	11.84	10.08/21.21/27.65	+0.11/+0.73/+0.85
`asym-region`	12.09	11.34/24.15/31.45	+0.08/+0.38/+0.53
`asym-image` (I)	11.40	10.96/24.13/31.52	+0.23/+0.43/+0.52
`asym-scene-part` (S)	12.24	12.27/26.43/33.91	+0.18/+0.28/+0.36

Table 8: **LoCaR cache storage** (MB). Combined denotes the sum of the reported gallery cache and support bank.

Backbone	Gallery cache	Support bank	Combined
CLIP-B/16	78	55	133
SigLIP2-B/16	116	80	196
Qwen3-VL-2B	270	28	298

$0.05/0.15/0.16$ pp. Thus, S improves region retrieval while leaving a smaller additional gain from paired support in this comparison. The matched global baselines differ from the standalone Qwen zero-shot R@1 of 11.78.

Image-cluster sign-flip tests give R@10 $p < 10^{-4}$ for model-default and object instructions, $p = 0.0001$ for region and image instructions, and $p = 0.0058$ for the scene-part instruction. The corresponding cluster-bootstrap 95% gain intervals are $[1.01, 1.58]$, $[0.57, 1.13]$, $[0.27, 0.80]$, $[0.28, 0.78]$, and $[0.11, 0.61]$ pp. These tests assess support gains within each instruction, not differences between instructions.

A.5 Cache storage

Table 8 reports the saved gallery-cache and support-bank sizes for each backbone. Their combined storage ranges from 133 to 298 MB. These figures describe stored files and do not measure peak memory consumption during scoring.

B Support-set contamination audit

Split-resolution failure. The public RefCOCOg mirror used in our earlier pipeline contains only `val-*` and `test-*` parquet shards. When requested to load the missing `train` split, a file-globbing fallback returns the union of validation and test. Separately, validation and test contain 481 shared rows with identical file and question identifiers. The declared split, bank size, and feature statistics remain plausible, so these checks alone do not establish support provenance.

Paired feature audit. Support anchors are raw features from the same frozen encoder used for evaluation. For each evaluation pair, we identify its nearest text anchor and nearest region anchor. We flag potential reuse only when both cosine similarities exceed τ and the two nearest-anchor indices agree. Requiring a shared index distinguishes paired overlap from repeated or formulaic language.

On the clean bank, lowering τ from 0.9999 to 0.95 increases text-only matches from 1.63% to 70.93%, while paired matches increase from 0.00% to 0.03%. The contaminated bank remains near 17% paired overlap throughout the sweep. Table 9 reports the strict-threshold results.

Effect on retrieval. The contaminated scorer achieves 59.3% R@1 on the 508 flagged queries, compared with 5.42% $[4.6, 6.4]$ on the remaining $2{,}492$. The much higher recall among flagged queries is consistent with support reuse inflating retrieval performance. A two-bank mixture calculation agrees on the clean rate but estimates a higher leaked rate because it assumes identical conditional behavior across banks. We retain that calculation as a cross-check and prioritize the direct

Table 9: **Support audit at $\tau = 0.9999$** on 3,000 RefCOCOg test queries. Paired overlap requires matching text and region features at the same support index. The 481 shared rows must be removed even when validation is loaded correctly.

Support bank	Declared source	Paired overlap
All shards via missing-split fallback	`train`	16.93%
Unfiltered validation	`val`	2.60%
Validation with shared test rows removed	`val`	0.00%

conditional measurements. RefCOCOg support banks are constructed from filtered validation data and must pass the audit before scoring.

C LocFG construction

Construction. From Visual Genome (Krishna et al., 2017) we emit two splits over shared images. *Retrieval*: a region description and the box it describes, at most 8 regions per image, phrases of at least three content words, deduplicated within an image. *Compositional*: image versus {correct, hard-negative} caption in three grounded categories — `attribute` (within-type attribute swap), `relation` (subject/object order swap of a directional predicate) and `object` (plausible absent same-frequency object replacement) — at most two items per category per image. Negatives are generated with length- and frequency-matching safeguards so a bag-of-words model cannot win by surface statistics; the length-bias check and a 50-per-category human-verification sample ship with the release.

Leakage control, which is the point. Images are filtered against RefCOCOg and the compositional benchmarks by a dHash plus banded-LSH blocklist, catching both COCO-identifier matches and perceptual near-duplicates. The blocklist is released alongside the data, because "leakage-controlled" is otherwise an unverifiable claim. `locfg_split.py` then carves an image-disjoint `fit`/`test` partition by a stable hash of the image identifier, so a method may be adapted in-domain on `fit` and evaluated on `test` without overlap.

A caveat we measured rather than inherited. Small referents are hard for class-agnostic proposers, and this bounds what any proposal-scoring method can do here. SAM2 at the settings TextRegion uses covers LocFG boxes poorly. Over all 12,033 evaluated queries, only 51.1% have a proposal reaching IoU 0.5 with the ground-truth box inside the $R_{\max}$=32 scoring cap (median 0.512), against 85.8% on RefCOCOg, and 16.8% have no proposal above IoU 0.1 at all. Lifting the cap recovers little — 54.2% at IoU 0.5 uncapped, median 0.549, 14.8% still unmatchable, saturating by a cap of 64 — so the referents are mostly not proposed, rather than proposed and then discarded. Users of LocFG who score proposals rather than ground-truth boxes should measure and report this; it is a property of the segmenter meeting small referents, not of the annotations.

D LocFG: eliminations in full

No γ rescues it, and the failure is domain-specific. Because our own calibration could not resolve γ , a null at one value would be weak evidence, so we swept the whole band and one value beyond it (Table 10). The transfer is null at every γ the support set left standing, and the trend runs *opposite* to RefCOCOg's. Beyond the band the two diverge sharply: at γ=0.30 the residual *costs* LocFG -1.94 R@10 (p=3.3×10^{-13}) while still *gaining* $+1.86$ R@10 on RefCOCOg (p=8.3×10^{-7}) with the identical bank and backbone. Upweighting the residual is therefore not generically harmful; it is harmful only where the exemplar structure does not apply, which is what an inert term would not do. The RefCOCOg column is the control that licenses this reading, and without it the out-of-band row would have been uninterpretable.

Coverage is necessary but not sufficient, and we tested that rather than assuming it. The obvious reconciliation of LocFG's flat profile is that its "well-covered" half is well covered only by its own standard — 51% of its queries clear IoU 0.5 against RefCOCOg's 86% (both over all

Table 10: **LocFG transfer, γ sweep**, RefCOCOg-built anchors applied unchanged. Δ against γ=0 on the region channel; 2,600 images / 12,033 queries. The RefCOCOg column is the same backbone and bank in-domain, at the same γ, as a control. Every LocFG entry inside the calibrated band $[0.05, 0.15]$ is non-significant; the out-of-band row separates the two datasets in *opposite* directions, both significantly.

	LocFG (transferred)				RefCOCOg
γ	ΔR@1	ΔR@5	ΔR@10	p (R@1 / R@10)	ΔR@10
0.05	+0.12	+0.02	+0.17	0.24 / 0.19	+1.24
0.10*	+0.08	+0.08	+0.12	0.59 / 0.50	+2.17
0.15	−0.03	−0.09	−0.16	0.88 / 0.45	+2.59
0.30 (out of band)	−0.61	−1.97	−1.94	$\mathbf{2.8\times10^{-3}}$ / $\mathbf{3.3\times10^{-13}}$	**+1.86**

evaluated queries at the $R_{\max}$=32 cap), its conditional anchor median 0.586 against 0.883 — so it may simply never reach the regime in which the depth gain exists. That predicts a depth gain on the LocFG queries that *do* reach RefCOCOg's regime. It is false. Splitting instead at IoU 0.8 isolates 2,542 LocFG queries whose referent SAM2 covers as well as a typical RefCOCOg one, and they show $+0.43/+0.12/+0.12$, with a top-minus-bottom contrast of +0.40, +0.11 and −0.22, none significant (p=0.31, 0.82, 0.65). Restoring coverage on LocFG does not restore the effect.

We therefore report coverage as *necessary and not sufficient*: within RefCOCOg, losing coverage removes the deep-rank gain; on LocFG, having coverage does not produce it. What distinguishes the two beyond coverage is not established here, and we flag one confound in our own test rather than leave it for a reader to find. The stratification controls coverage on the *query* side only; the LocFG bank itself was built from anchors with median IoU 0.586, so even a well-covered query is matched against noisy exemplars. We tested that: rebuilding the bank at `-iou-min 0.7` yields 2,827 anchors at median IoU 0.836, comparable to RefCOCOg's 0.883, and the aggregate does not improve (−0.02 R@1; the small decline is within what the scaling curve attributes to the smaller bank). Bank quality is therefore not the obstacle either.

What replicates across both banks is narrower and stranger. The R@1 coverage contrast is +0.76 (p=8.2 × 10^{-3}) with the unfiltered bank and +0.75 (p=8.4 × 10^{-3}) with the filtered one — two independently constructed banks agreeing to the second decimal — while the R@10 contrast is +0.30 and +0.06, neither significant. LocFG thus has a small, coverage-gated rank-1 effect and no depth effect at all, which is the mirror image of RefCOCOg, where coverage gates the depth effect (+1.89, p=0.021) and leaves rank-1 untouched (−0.23, p=0.71). We do not have an account of why the two invert, and we prefer to report the inversion than to supply one after the fact; two of our earlier explanations for this benchmark were falsified by the next control we ran.

A third candidate is also eliminated by measurement: that string-unique VG phrases describe visually *commonplace* referents, so deep-rank promotion is spread across lookalike regions in other images and nets nothing. Counting gallery images that contain a near-copy of each query's referent (region-feature cosine ≥ 0.9) gives a *median of zero* on LocFG; the unique-vs-most-duplicated contrast is flat at every level (R@10 +0.34, p=0.58); and RefCOCOg in fact contains *more* such duplication (median 4 vs 2 at cosine ≥ 0.85) while keeping its depth gains. Whatever suppresses the depth component here, it is not that the referents are commonplace, and it is not proposal coverage, and it is not bank quality.

The last candidate we could name was the query distribution itself: RefCOCOg expressions are *discriminative* by construction, Visual Genome descriptions merely describe, and a mechanism that matches paired exemplars might need the contrastive structure. The one operational form we could derive from that conjecture is profile collapse — descriptive queries selecting the same few anchors, driving the residual toward the query-*independent* per-image score that the region-marginal control already measures as inert. Measured, the collapse is absent, and absent on both datasets alike: queries use 1,676 of 4,096 anchors effectively on LocFG against 1,923 on RefCOCOg, mean pairwise profile cosine is 0.008 on both, and the mean correlation between residual score rows is 0.037 against 0.034. The residual is as query-dependent on the benchmark where it does nothing as on the benchmark where it works.

An eval-read ledger for this benchmark. The experiments above read LocFG test under six scoring configurations: the pre-registered transfer (γ=0.10, one read), three band-sensitivity values ($\gamma \in \{0.05, 0.15, 0.30\}$, labelled analysis), the in-domain bank at transferred hyperparameters, and the IoU-filtered in-domain bank. The stratifications (coverage at two thresholds, and the diagnostics below) re-partition those same scoring passes rather than scoring new configurations, and select nothing. We state this because the count is high enough that a reader should get to audit it rather than trust it: none of the six was chosen on the basis of a previous read's outcome except as disclosed retractions in this section, and any future configuration on this benchmark — a support-selected region cap, a well-posedness query filter, an in-domain calibration — will be frozen on the fit split and read exactly once, extending this ledger.