GRANULARITY-AWARE EVIDENCE AGGREGATION AND CALIBRATION FOR TRAINING-FREE ZERO-SHOT MULTI-LABEL RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

CLIP provides strong zero-shot recognition capabilities, but its global representation may overlook small or spatially localized objects in multi-label images. Local crops can recover such evidence, yet may also amplify responses to absent classes, including semantically related alternatives. We introduce GRADE (Granularity-Aware Regional Aggregation and Discriminative Evidence Calibration), a training-free framework for zero-shot multi-label image recognition. GRADE aggregates class-specific regional responses across multiple context granularities, using class-wise global-score percentiles to guide aggregation. When local evidence dominates, comparison with semantically neighboring classes downweights ambiguous local responses before global-local fusion. Experiments on COCO and VG-256 demonstrate competitive performance against existing training-free methods, with consistent mAP improvements over global CLIP across both benchmarks. Ablation studies further confirm the effectiveness of aggregating local evidence across multiple regions and context granularities.

1 INTRODUCTION

Vision-language models (VLMs), such as CLIP (Radford et al., 2021), have shown strong zero-shot recognition ability by aligning images and text in a shared embedding space. This capability makes them well suited to multi-label image recognition (MLR), where multiple visual concepts may coexist in a single image (Zhang & Zhou, 2014). However, standard CLIP inference relies on a single global image representation, which tends to emphasize dominant and salient objects. As a result, CLIP can overlook small or less prominent objects (Lin et al., 2024).

A natural way to address this limitation is to exploit local visual evidence. DenseCLIP shows that image–text matching can support dense, localized prediction (Rao et al., 2022). For zero-shot MLR, TagCLIP shows that patch- or region-level representations can reveal objects weakly captured by the global image representation (Lin et al., 2024), while PIAA demonstrates that patch-level scores can be aggregated adaptively (Wang et al., 2026). However, local evidence is not always reliable. In the top example of Fig. 1, the global scores rank the true class *car* 49th, whereas the local crop ranks it first. In the bottom example, a local crop dominated by hair texture raises the absent class *hair drier* from global rank 79 to local rank 1. Local crops can thus produce strong responses to absent classes, a phenomenon we refer to here as *local hallucination*. This term describes elevated class responses rather than errors under a specified binary decision threshold. These examples highlight the dual role of local evidence: it can recover overlooked classes, but can also amplify misleading visual cues. A key challenge is therefore to exploit informative local evidence while preventing unreliable local responses from overwhelming global predictions.

Unlike prompt-learning approaches that optimize continuous context (Zhou et al., 2022), our setting excludes target-task annotations and parameter optimization. A single region or fixed context granularity may miss relevant evidence, whereas broader regional coverage can also introduce ambiguous local responses. This motivates aggregating evidence across regions and context granularities while controlling its contribution to the final scores.

Motivated by this observation, we investigate how local evidence can be used more effectively for CLIP-based zero-shot multi-label recognition. To this end, we introduce **GRADE** (**G**ranularity-

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.

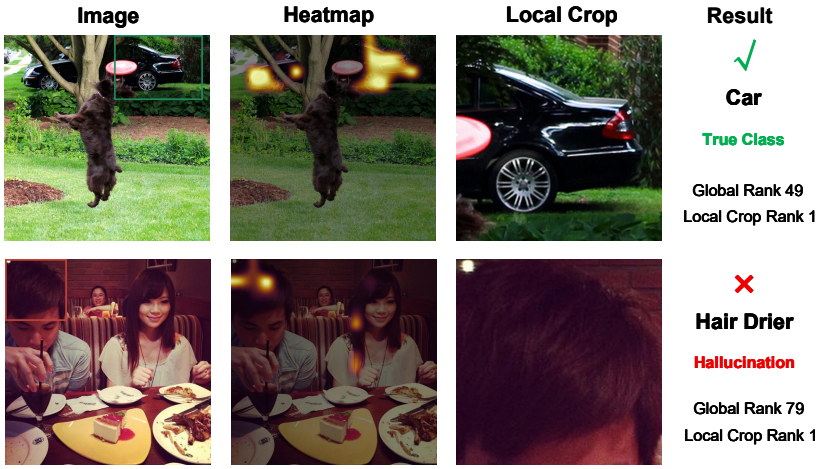


Figure 1: Local evidence can recover overlooked classes but also amplify responses to absent classes. In the top example, the true class *car* moves from global rank 49 to local rank 1. In the bottom example, the absent class *hair drier* moves from global rank 79 to local rank 1, illustrating a local hallucination through its elevated class response.

Aware **R**egional **A**ggregation and **D**iscriminative **E**vidence Calibration), a training-free framework that improves regional coverage and selectively calibrates local evidence. GRADE first derives class-specific regions from CLIP heatmaps and aggregates their responses across multiple regions and granularities to improve evidence coverage. We then estimate class-wise global score distributions from a small unlabeled subset of the target training pool and use the resulting percentiles to calibrate the aggregation of multi-granularity local responses. For predictions dominated by local evidence, a semantic consistency mechanism compares the selected region against semantically related classes and suppresses responses that lack discriminative support. The entire framework requires neither parameter optimization nor target-task annotations; the reference images are used only to estimate score distributions. Our main contributions are summarized as follows:

- We highlight the dual role of local evidence in CLIP-based multi-label recognition: it can recover classes overlooked by global representations, but can also produce local hallucinations through strong responses to absent classes. This observation motivates jointly addressing evidence coverage and reliability.
- We propose training-free multi-region, multi-granularity aggregation, with distribution-based and semantic checks that selectively adjust ambiguous local responses.
- Experiments on two benchmarks show consistent mAP gains over global CLIP. Ablations attribute most of the improvement to regional aggregation, with smaller, dataset-dependent effects from calibration.

2 Related Work

Multi-Label Image Recognition. Multi-label recognition must resolve co-occurring categories and their dependencies. Feature-selection methods exploit feature correlations (Yin et al., 2024) and semantic label relations (Chen et al., 2025); discriminative structure learning (Jia et al., 2024) and contrastive label enhancement (Chen et al., 2027) further refine label representations. Distributional supervision models relative label importance (Zhang et al., 2021), while imbalance-aware selection addresses uneven class frequencies (Liu et al., 2025). Image models capture dependencies through graph convolution (Chen et al., 2019), spatial attention (Zhu et al., 2017), Transformers (Lanchantin et al., 2021), or imbalance-aware objectives (Ridnik et al., 2021). These approaches rely on target labels; GRADE studies training-free inference without them.

CLIP-Based Multi-Label Recognition. CLIP enables zero-shot recognition through image–text matching (Radford et al., 2021). Prompt adaptation includes positive and negative prompts (Sun et al., 2022) and text-derived prompts (Guo et al., 2023). Zero-shot alternatives learn diverse semantic embeddings (Ben-Cohen et al., 2021), transfer multimodal knowledge (He et al., 2023), or train language-derived cross-modal classifiers (Liu et al., 2024). Label distributions also support semantic discrimination in related cross-modal retrieval (Qi et al., 2026). Training-free methods include TagCLIP's local-to-global evidence (Lin et al., 2024), SPARC's score fusion (Miller et al., 2025), and DualPrompt's co-occurrence prompts (Xie et al., 2026).

Regional Evidence Aggregation and Calibration. TagCLIP extracts patch-level evidence (Lin et al., 2024), PIAA adaptively aggregates patch responses (Wang et al., 2026), and SPARC standardizes and fuses image–text scores (Miller et al., 2025). GRADE instead aggregates distinct class-specific connected regions across crop contexts, using class-wise global-score ranks to select among context summaries. Its semantic check applies only when the selected local response dominates the global score.

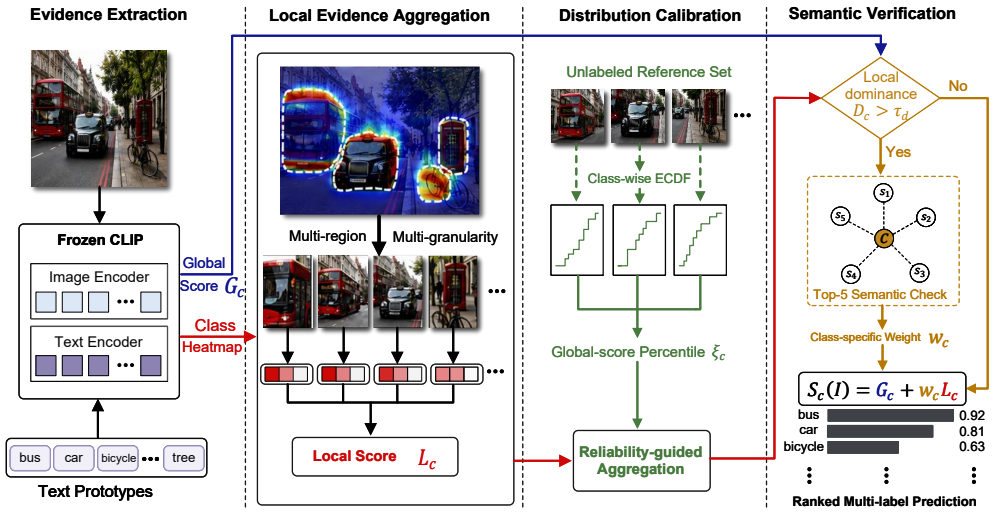


Figure 2: Overview of GRADE. Class-specific regional responses are aggregated across context granularities using global-score percentiles. The strongest calibrated region undergoes semantic verification when local evidence dominates, before fusion with the global score.

3 Preliminaries

3.1 Problem Formulation

Given an image I and a target vocabulary $\mathcal{C} = \{c_1, \ldots, c_C\}$, zero-shot MLR assigns a relevance score to each class without labeled examples from the target task. The annotation $\mathbf{y} \in \{0,1\}^C$ is used only for evaluation, with $y_c = 1$ indicating that class c is present. The model returns

$$\mathbf{s}(I) = [s_1(I), \ldots, s_C(I)]^\top \in \mathbb{R}^C, \tag{1}$$

where $s_c(I)$ measures the relevance of class c to I. These scores are evaluated using ranking metrics and the thresholding protocol specified in Section 5.1. The notation used throughout the paper is summarized in Appendix A.

3.2 CLIP-based Zero-shot Multi-label Recognition

CLIP uses a visual encoder E_{v} and a text encoder E_{t} to align images and text in a shared space (Radford et al., 2021). We keep both encoders frozen. For class c, an ensemble of M prompt

templates gives the text prototype

$$\mathbf{t}_c = \operatorname{norm}\left(\frac{1}{M}\sum_{m=1}^{M}\operatorname{norm}\left(E_{\mathrm{t}}(p_m(c))\right)\right), \tag{2}$$

where $\operatorname{norm}(\mathbf{z}) = \mathbf{z}/\max(\|\mathbf{z}\|_2, \epsilon)$ with $\epsilon = 10^{-12}$ for numerical stability. For an input image, the normalized global feature, global logits, and class scores are

$$\begin{aligned} \mathbf{g} &= \operatorname{norm}\left(E_{\mathrm{v}}(I)\right), \qquad a_c^G = \alpha\,\mathbf{g}^\top \mathbf{t}_c, \\ G_c &= \frac{\exp(a_c^G)}{\sum_{k=1}^{C}\exp(a_k^G)}, \qquad c = 1, \ldots, C, \end{aligned} \tag{3}$$

where α is the pretrained CLIP logit scale, which remains frozen; the cached runs use $\alpha = 100$. With a ViT visual encoder (Dosovitskiy et al., 2021), $\mathbf{g}$ is obtained from the final class token. We write $G_c(I)$ when the image argument must be explicit. We use vocabulary-normalized similarities as relevance scores, not as independent class-presence probabilities.

3.3 Dense patch features from CLIP

The global token can underrepresent small or localized objects. Following TagCLIP (Lin et al., 2024), we derive local tokens from the penultimate-layer sequence using a modified final Transformer block (Vaswani et al., 2017). In the local branch, an additive attention mask has zero diagonal entries and -100 off-diagonal entries, suppressing cross-token attention while retaining the attention projections, residual connections, LayerNorm, and MLP. The class token remains in the masked sequence but is discarded when extracting patch tokens; the global branch uses the unmodified final block. Earlier blocks are unchanged, so local tokens retain information mixed in preceding layers. The resulting token $\mathbf{z}_i^{\mathrm{loc}}$ is mapped to the joint embedding space by

$$\mathbf{u}_i = \operatorname{norm}\left(\mathrm{LN}_{\mathrm{post}}(\mathbf{z}_i^{\mathrm{loc}})\mathbf{W}_{\mathrm{p}}\right), \qquad i = 1, \ldots, N. \tag{4}$$

Using the same vocabulary normalization as the global branch, the response of patch i to class c is

$$P_{i,c} = \frac{\exp\left(\alpha\,\mathbf{u}_i^\top \mathbf{t}_c\right)}{\sum_{k=1}^{C}\exp\left(\alpha\,\mathbf{u}_i^\top \mathbf{t}_k\right)}. \tag{5}$$

Arranging $\{P_{i,c}\}_{i=1}^{N}$ spatially yields the raw class map $\mathbf{H}_c$. Let $h_c^{\min} = \min(\mathbf{H}_c)$ and $h_c^{\max} = \max(\mathbf{H}_c)$. Its normalized form is

$$\widehat{\mathbf{H}}_c = \begin{cases} \dfrac{\mathbf{H}_c - h_c^{\min}}{h_c^{\max} - h_c^{\min}}, & h_c^{\max} - h_c^{\min} > \epsilon, \\ \mathbf{0}, & \text{otherwise}, \end{cases} \tag{6}$$

which provides the class-specific spatial evidence used by GRADE.

4 Method

GRADE uses the global scores and class-specific response maps defined in Section 3 to extract multiple regions and evaluate them across context granularities. Global-score percentiles guide the aggregation of each region's responses, after which the strongest calibrated region is selected. When local evidence dominates, semantic verification adjusts its contribution before global-local fusion.

4.1 Granularity-aware regional aggregation

Class-Specific Region Discovery. The normalized map identifies where a class response occurs, but a single peak can miss spatially separated evidence. We retain classes whose maximum patch response in Eq. (5) exceeds an image-specific threshold. The candidate set is

$$\mathcal{C}_I = \left\{ c \in \mathcal{C} : \max_{1 \le i \le N} P_{i,c} > P_{\min} + \tau_{\mathrm{cand}}(P_{\max} - P_{\min}) \right\}, \tag{7}$$

where $P_{\min} = \min_{i,k} P_{i,k}$ and $P_{\max} = \max_{i,k} P_{i,k}$. For each $c \in \mathcal{C}_I$, we threshold $\widehat{\mathbf{H}}_c$ at η and extract connected components. After sorting them by area, we retain up to R components and use their bounding boxes as the class-specific regions $\{\mathcal{B}_{c,r}\}_{r=1}^{R_c}$, where $R_c \leq R$. Multiple components preserve distinct local hypotheses; the amount of context assigned to each hypothesis is determined next.

Multi-Granularity Regional Evidence (MGE). A detected box localizes evidence but does not specify how much surrounding context is useful. A tight crop preserves object detail, whereas a broader crop may recover missing context or introduce distractors. We therefore construct J crops around each $\mathcal{B}_{c,r}$ using relative margins $\{\rho_j\}_{j=1}^{J}$, expanding each horizontal side by ρ_j times the box width and each vertical side by ρ_j times its height, then clipping to the image boundaries. All crops are square-padded and resampled to the same encoder resolution. Granularity denotes the crop's context extent; resizing also changes the object's relative scale. Appendix C specifies rounding and padding. Let $I_{c,r}^{(j)}$ be the resulting crop and $\mathbf{v}_{c,r}^{(j)} = \mathrm{norm}(E_{\mathrm{v}}(I_{c,r}^{(j)}))$ its visual representation. Using the same class-wise normalization as Eq. (3), its response to class c is

$$\ell_{c,r}^{(j)} = \frac{\exp\left(\alpha(\mathbf{v}_{c,r}^{(j)})^\top \mathbf{t}_c\right)}{\sum_{k=1}^{C} \exp\left(\alpha(\mathbf{v}_{c,r}^{(j)})^\top \mathbf{t}_k\right)}. \tag{8}$$

We summarize the responses of each region by

$$\ell_{c,r}^{\min} = \min_j \ell_{c,r}^{(j)}, \qquad \ell_{c,r}^{\mathrm{geo}} = \left(\prod_{j=1}^{J} \ell_{c,r}^{(j)}\right)^{1/J}, \qquad \ell_{c,r}^{\max} = \max_j \ell_{c,r}^{(j)}. \tag{9}$$

The response sequence across j records how the evidence changes with context. The minimum and maximum provide conservative and permissive summaries, respectively, while the geometric mean favors responses that remain strong across granularities. Stable evidence keeps these summaries close, whereas a context-sensitive response separates them. Rather than fixing one summary for every image and class, we select along this range using the global evidence.

4.2 Discriminative evidence calibration

Expanding regions and context improves coverage, but it also increases the risk of two types of errors. A response may occur only at a favorable granularity, or it may be driven by visual cues shared with a related class. GRADE addresses these two issues in sequence. Distribution guidance first determines the required degree of agreement across granularities, while semantic verification subsequently examines the selected regional evidence.

Distribution-Guided Granularity Calibration (DGC). We use more conservative aggregation when the global score is low relative to the class-specific reference distribution. Because the scale of G_c differs across classes, we estimate its relative position using an unlabeled reference subset $\mathcal{D}_{\mathrm{ref}} = \{I_n^{\mathrm{ref}}\}_{n=1}^{N_{\mathrm{ref}}}$, where $N_{\mathrm{ref}} > 1$. For a query image I, let $k_c(I)$ count reference scores not exceeding $G_c(I)$. The standard ECDF is $k_c(I)/N_{\mathrm{ref}}$; we instead use the following ECDF-based normalized rank as our percentile convention:

$$\xi_c(I) = \mathrm{clip}_{[0,1]}\left(\frac{\sum_{n=1}^{N_{\mathrm{ref}}} \mathbb{I}[G_c(I_n^{\mathrm{ref}}) \leq G_c(I)] - 1}{N_{\mathrm{ref}} - 1}\right). \tag{10}$$

Eq. (10) equals $\mathrm{clip}_{[0,1]}((k_c(I) - 1)/(N_{\mathrm{ref}} - 1))$. Ties use the right cumulative count: all reference scores equal to the query are included. Scores below the minimum map to zero, and scores at or above the maximum map to one; ties at the minimum need not map to zero. This is a rank rescaling, not a leave-one-out correction. The normalized rank measures relative global support, not the probability that the class is present. Omitting its argument I below, we calibrate each regional response across granularities as

$$\widetilde{\ell}_{c,r} = \begin{cases} (1 - 2\xi_c)\ell_{c,r}^{\min} + 2\xi_c \ell_{c,r}^{\mathrm{geo}}, & \xi_c \leq \frac{1}{2}, \\ 2(1 - \xi_c)\ell_{c,r}^{\mathrm{geo}} + (2\xi_c - 1)\ell_{c,r}^{\max}, & \xi_c > \frac{1}{2}. \end{cases} \tag{11}$$

When $\xi_c \leq \frac{1}{2}$, the global evidence for class c is relatively weak, and the aggregation is therefore biased toward cross-granularity consistency by interpolating between $\ell_{c,r}^{\min}$ and $\ell_{c,r}^{\text{geo}}$. Otherwise, stronger global support allows more reliance on the strongest local response, and the aggregation interpolates between $\ell_{c,r}^{\text{geo}}$ and $\ell_{c,r}^{\max}$. In this way, the required degree of agreement across granularities is relaxed progressively as the class-relative global support increases. We then select the strongest calibrated region:

$$L_c = \max_{1 \leq r \leq R_c} \widetilde{\ell}_{c,r}, \qquad r_c^\star = \arg\max_{1 \leq r \leq R_c} \widetilde{\ell}_{c,r}. \tag{12}$$

We set $L_c = 0$ when $c \notin \mathcal{C}_I$ or no component is retained. Only $r_c^\star$ contributes to L_c, so the subsequent semantic check operates on this selected region rather than all candidates.

Semantic Consistency Verification (SCV). Distribution guidance controls sensitivity to context, but it cannot rule out semantic interference. A region may remain strong across granularities because its appearance is also compatible with a related class. We examine this residual ambiguity only when local evidence dominates:

$$D_c = \frac{L_c}{\max(G_c + L_c, \epsilon)}. \tag{13}$$

For $D_c \geq \tau_d$, let $\mathcal{N}_K(c)$ contain the K classes in $\mathcal{C} \setminus \{c\}$ with the highest text-prototype cosine similarity to c, explicitly excluding the target class itself. For each $k \in \mathcal{N}_K(c)$, we encode the paired prompts "a photo of c rather than k" and "a photo of k rather than c" as normalized embeddings $\mathbf{e}_{c \succ k}$ and $\mathbf{e}_{k \succ c}$. Their difference defines the discriminative direction

$$\mathbf{d}_{c,k} = \text{norm}\left(\mathbf{e}_{c \succ k} - \mathbf{e}_{k \succ c}\right). \tag{14}$$

For the region $r_c^\star$ selected in Eq. (12), we compute the target-versus-neighbor margin and semantic support rate

$$q_{c,k} = \frac{1}{J} \sum_{j=1}^{J} \left(\mathbf{v}_{c,r_c^\star}^{(j)}\right)^\top \mathbf{d}_{c,k}, \qquad W_c = \frac{1}{K} \sum_{k \in \mathcal{N}_K(c)} \mathbb{I}[q_{c,k} > 0]. \tag{15}$$

A positive $q_{c,k}$ indicates that the selected region supports c over neighbor k. We convert the fraction of positive comparisons into a reliability weight:

$$w_c = \begin{cases} \min(1, 2W_c), & D_c \geq \tau_d, \\ 1, & D_c < \tau_d. \end{cases} \tag{16}$$

For locally dominant responses, SCV attenuates the local contribution when fewer than half of the pairwise comparisons favor the target, i.e., $D_c \geq \tau_d$ and $W_c < 0.5$. Otherwise, $w_c = 1$; this does not require every neighbor comparison to favor the target.

4.3 Global–local evidence fusion

Using the score notation in Eq. (1), the final prediction for class c after granularity calibration and semantic verification is

$$s_c(I) = G_c + w_c L_c. \tag{17}$$

The resulting vector $\mathbf{s}(I)$ contains the final class scores. SCV attenuates the local increment while preserving the global score: $s_c(I) \geq G_c$ because $L_c \geq 0$ and $w_c \in [0, 1]$. When no local region is available, $L_c = 0$ and the score reduces to G_c. GRADE introduces no trainable module or parameter updates; reference images are used only to compute the normalized ranks in Eq. (10), and their annotations do not enter score estimation or prediction.

5 Experiments

5.1 Datasets and Evaluation Metrics

We evaluate on COCO (Lin et al., 2014) and VG-256 (Krishna et al., 2017), with 80 and 256 classes and 40,137 and 32,011 evaluation images in the aligned splits, respectively (Appendix B). We report threshold-free mean average precision (mAP) and macro-averaged F1 (Macro-F1). Macro-F1 uses per-class thresholds estimated from the evaluation-score distribution at a fixed 2% predicted-positive rate; it describes that operating point, not a deployment threshold (Appendix C). P@1, P@3, and P@5 are reserved for ablations. All values are percentages.

5.2 Experimental Settings

All methods use CLIP ViT-B/16 at 224×224; only PIAA_SC uses SC-CLIP (Bai et al., 2025) patch extraction. GRADE freezes CLIP and uses up to five regions, margins $\{0.1, 0.2, 0.4\}$, and 1% unlabeled reference images. On COCO, this entails 64.51 regional crop evaluations per image on average; this count is not a latency measurement (Appendix C). DualCoOp trains on the corresponding labeled images and selects checkpoints on evaluation data. Appendices C and D give protocol and budget details. The GRADE experiments were run on an NVIDIA RTX 4090 GPU. Anonymous GRADE code and reproduction instructions are available at `https://anonymous.4open.science/r/GRADE2026`.

5.3 Comparison Methods

We compare with global CLIP (G), project-specific DualCoOp-TF, TagCLIP (Lin et al., 2024), SPARC, trained DualCoOp (Sun et al., 2022), PIAA (Wang et al., 2026), PIAA_SC, and DualPrompt (Xie et al., 2026) under the protocols in Appendix C. Table 1 distinguishes *Ours*, the global-plus-regional variant without DGC and SCV, from complete GRADE.

Table 1: Main comparison (%). Bold marks the best within the None and 1% training/reference-data blocks for each metric. Both SPARC rows jointly calibrate on the unlabeled evaluation split. *Ours* omits DGC and SCV; GRADE is complete.

Method	Target-data access	Backbone	Resolution	COCO		VG-256	
				mAP	Macro-F1	mAP	Macro-F1
CLIP	None	ViT-B/16	224×224	65.05	55.87	30.47	29.27
DualCoOp-TF	None	ViT-B/16	224×224	34.93	33.03	17.54	18.09
TagCLIP	None	ViT-B/16	224×224	71.02	59.19	36.65	33.60
SPARC	None	ViT-B/16	224×224	66.17	56.17	34.88	32.46
Ours	None	ViT-B/16	224×224	**72.15**	**60.71**	**37.36**	**34.53**
DualCoOp	1% labeled	ViT-B/16	224×224	54.37	47.87	26.87	26.34
PIAA	1% unlabeled	ViT-B/16	224×224	66.92	56.72	31.28	29.76
PIAA_SC	1% unlabeled	ViT-B/16	224×224	69.38	58.63	34.91	31.98
DualPrompt	1% labeled	ViT-B/16	224×224	69.36	58.51	36.13	33.67
SPARC	1% labeled	ViT-B/16	224×224	67.12	56.72	36.11	33.36
GRADE (Ours)	1% unlabeled	ViT-B/16	224×224	**72.62**	**61.10**	**37.63**	**34.76**

5.4 Results and Analysis

Table 1 shows that GRADE improves mAP over frozen global CLIP by 7.57 and 7.16 percentage points (pp) on COCO and VG-256, respectively; Macro-F1 increases by 5.23 and 5.49 pp.

With no training or reference images, *Ours* leads on both datasets. SPARC's normalization and PCA use the entire unlabeled evaluation split jointly, whereas *Ours* scores images independently. Selecting SPARC's pair prompts from 1% labeled training data raises its mAP and Macro-F1 on both datasets. In the 1% block, GRADE leads on COCO and VG-256 despite using unlabeled data. DGC and SCV add only 0.47/0.27 pp mAP to *Ours*, indicating that regional aggregation supplies most of GRADE's improvement.

5.5 Qualitative Analysis

We use two selected COCO cases to separate coverage from reliability. In Fig. 3, the useful *traffic light* region is third by component area and would be lost with one region, whereas a railway lamp elicits *parking meter* at every granularity. Neither a strong peak nor persistence across crops therefore establishes class identity. Fig. 4 shows that DGC preserves the class-relatively strong traffic-light evidence but treats parking meter conservatively. Because both predictions remain locally dominated, SCV compares semantic preference, retaining the true contribution and removing the false one without altering G_c. These selected cases illustrate the two decisions, not their frequency.

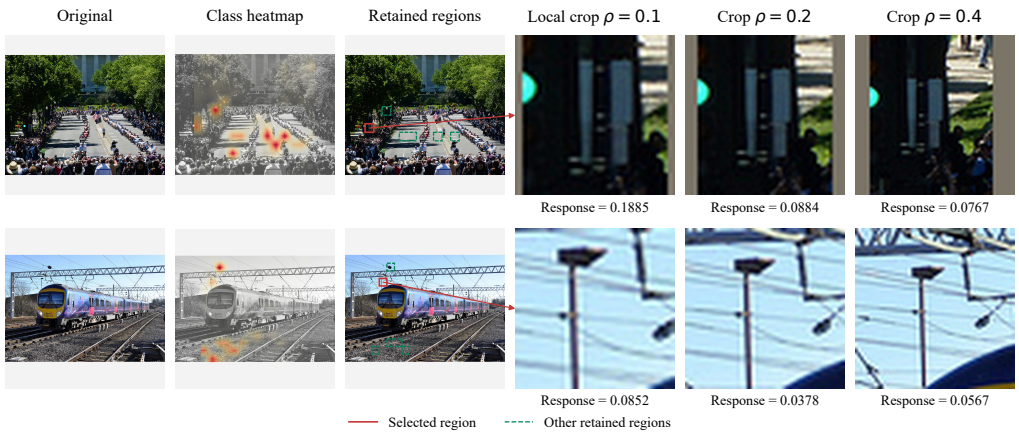
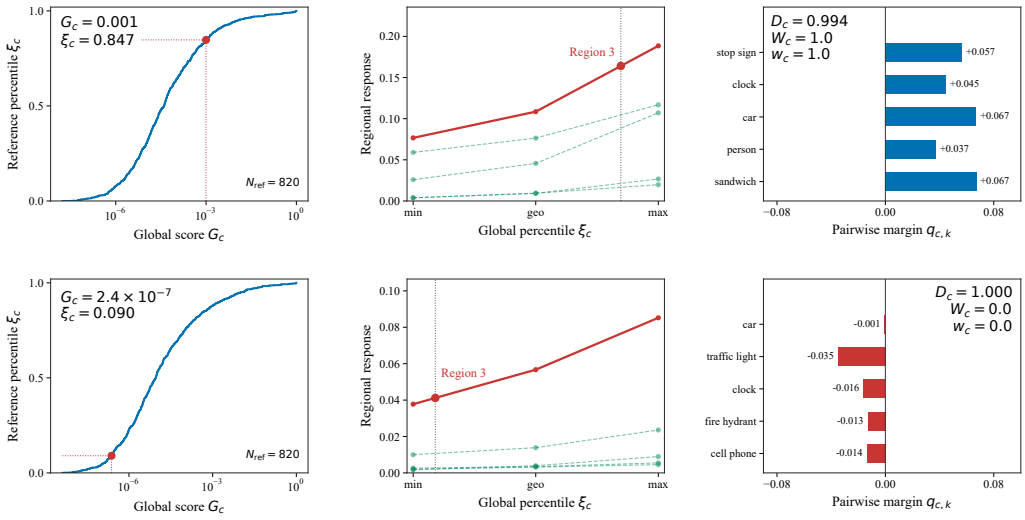


Figure 3: COCO regional evidence for image IDs 565635 (top) and 277955 (bottom): *traffic light* is present and *parking meter* is absent, respectively. Heatmaps show normalized responses; red marks the region selected after DGC and green the alternatives. Arrows link selected regions to fixed-resolution crops.



(a) Global percentile (b) DGC (c) SCV

Figure 4: ECDF, DGC, and SCV for Fig. 3, in the same row order. ECDFs use 820 unlabeled images; min/geo/max correspond to $\xi_c = 0/0.5/1$.

5.6 Ablation Study

Table 2 reports cumulative ablations using the same frozen-CLIP cache and evaluation split for each dataset. Starting from G, we successively add one crop, MGE, additional regions, DGC, and SCV. Only DGC uses the 1% unlabeled reference subset. The fusion coefficient remains fixed at one, while SCV applies the class-specific weight w_c. P@k measures precision among each image's top-k labels. Deltas use unrounded scores, and table entries are rounded to two decimals.

Table 2: Cumulative ablations (%). Every row retains G. L enables local evidence and R caps regions. ✓/– indicate presence/absence. Bold marks the best per dataset and metric, including rounded ties.

Dataset	L	MGE	R	DGC	SCV	mAP	P@1	P@3	P@5
COCO	–	–	–	–	–	65.05	88.55	46.94	33.82
	✓	–	1	–	–	70.81	91.25	52.61	37.19
	✓	✓	1	–	–	71.36	91.46	52.88	37.44
	✓	✓	5	–	–	72.15	91.56	53.20	37.72
	✓	✓	5	✓	–	72.56	91.61	53.06	37.78
	✓	✓	5	✓	✓	**72.62**	**91.66**	**53.22**	**37.85**
VG-256	–	–	–	–	–	30.47	68.77	46.29	37.28
	✓	–	1	–	–	36.43	75.47	56.81	46.64
	✓	✓	1	–	–	36.80	75.06	56.81	46.85
	✓	✓	5	–	–	37.36	75.58	58.37	48.14
	✓	✓	5	✓	–	**37.68**	76.01	58.81	**48.50**
	✓	✓	5	✓	✓	37.63	**76.03**	**58.83**	48.49

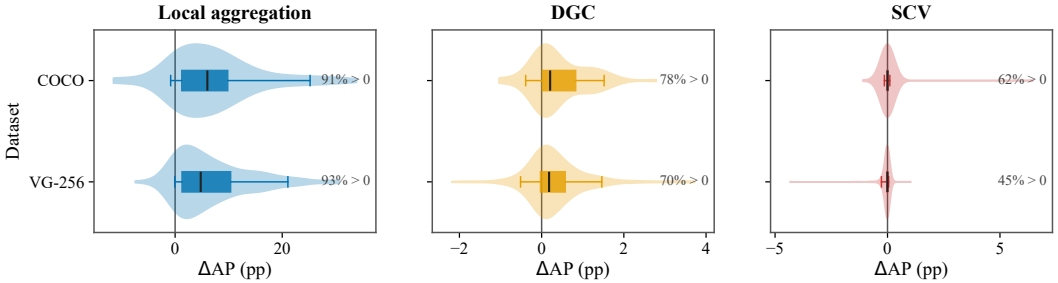


Figure 5: Class-wise ΔAP for local aggregation over G, DGC over local aggregation, and SCV over DGC. Boxes show interquartile ranges. Annotations give fractions of classes with positive change.

Regional aggregation without calibration accounts for 93.8–96.2% of GRADE's mAP gain over G. With one region, MGE improves mAP by 0.55 and 0.37 pp on COCO and VG-256, respectively, although VG-256 P@1 decreases. Increasing R from one to five improves all four metrics on both datasets, indicating that context diversity and regional coverage provide complementary benefits. Fig. 5 further shows that regional aggregation improves AP for more than 90% of classes in both datasets. Calibration yields smaller and less uniform changes. DGC adds 0.408 and 0.322 pp mAP on COCO and VG-256, respectively. SCV subsequently changes mAP by $+0.063$ and -0.052 pp, while its class-wise effects cluster near zero. These results support semantic preference as a selective, fallible reliability proxy whose effect is not uniformly positive. The cumulative comparisons measure effects conditional on preceding stages, rather than independent component contributions. Appendices D.2–D.3 provide class-wise and COCO diagnostic analyses.

6 Conclusion

GRADE combines regional evidence across context granularities with distribution-guided calibration and semantic verification. It improves mAP over global CLIP on both benchmarks without target-task labels or parameter updates. Regional aggregation contributes most of the gain, while calibration is smaller and dataset-dependent; SCV slightly reduces VG-256 mAP. Additional crop evaluations and reliance on unlabeled reference images remain practical considerations. The method also inherits the limits of its semantic verifier, whose effect is not uniformly positive across datasets. Future work could reduce crop cost through adaptive region budgets, estimate calibration statistics with less target-domain data, and develop uncertainty-aware verification while preserving the training-free setting.

REFERENCES

Sule Bai, Yong Liu, Yifei Han, Haoji Zhang, Yansong Tang, Jie Zhou, and Jiwen Lu. Self-calibrated CLIP for training-free open-vocabulary segmentation. *IEEE Trans. Image Process.*, 34:8271–8284, 2025.

Avi Ben-Cohen, Nadav Zamir, Emanuel Ben Baruch, Itamar Friedman, and Lihi Zelnik-Manor. Semantic diversity learning for zero-shot multi-label classification. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 640–650, 2021.

Liangzhou Chen, Mingjie Cai, and Qingguo Li. Multi-label feature selection with high-level semantic label relationships based on fuzzy rough sets. *Fuzzy Sets and Systems*, 510:109368, 2025. doi: 10.1016/j.fss.2025.109368.

Liangzhou Chen, Mingjie Cai, and Chaoqun Huang. Contrastive label enhancement-based multi-label feature selection. *Pattern Recognition*, 183:114879, 2027. doi: 10.1016/j.patcog.2026.114879.

Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 5177–5186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

Zixian Guo, Bowen Dong, Zhilong Ji, Jinfeng Bai, Yiwen Guo, and Wangmeng Zuo. Texts as images in prompt tuning for multi-label image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 2808–2817, 2023.

Sunan He, Taian Guo, Tao Dai, Ruizhi Qiao, Xiujun Shu, Bo Ren, and Shu-Tao Xia. Open-vocabulary multi-label classification via multi-modal knowledge transfer. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 808–816, 2023.

Qingwei Jia, Tingquan Deng, Yan Wang, and Changzhong Wang. Discriminative label correlation based robust structure learning for multi-label feature selection. *Pattern Recognit.*, 154:110583, 2024.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017.

Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 16478–16488, 2021.

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 *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755, 2014.

Yuqi Lin, Minghao Chen, Kaipeng Zhang, Hengjia Li, Mingming Li, Zheng Yang, Dongqin Lv, Binbin Lin, Haifeng Liu, and Deng Cai. TagCLIP: A local-to-global framework to enhance open-vocabulary multi-label classification of CLIP without training. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI*, pp. 3513–3521, 2024.

Qiong Liu, Mingjie Cai, Qingguo Li, and Chaoqun Huang. Multi-label feature selection based on adaptive label enhancement and class-imbalance-aware fuzzy information entropy. *International Journal of Approximate Reasoning*, 176:109320, 2025. doi: 10.1016/j.ijar.2024.109320.

Yicheng Liu, Jie Wen, Chengliang Liu, Xiaozhao Fang, Zuoyong Li, Yong Xu, and Zheng Zhang. Language-driven cross-modal classifier for zero-shot multi-label image recognition. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32173–32183, 2024.

Kevin Miller, Aditya Gangrade, Samarth Mishra, Kate Saenko, and Venkatesh Saligrama. SPARC: score prompting and adaptive fusion for zero-shot multi-label recognition in vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 4313–4321, 2025.

Gongao Qi, Xiangnan Zhou, and Chaoqun Huang. Label distribution-driven semantic discrimination enhanced hashing for cross-modal retrieval. *Information Processing & Management*, 63(7): 104853, 2026. doi: 10.1016/j.ipm.2026.104853.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. DenseCLIP: Language-guided dense prediction with context-aware prompting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 18082–18091, 2022.

Tal Ridnik, Emanuel Ben Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 82–91, 2021.

Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017.

Akang Wang, Xili Deng, Zhanxuan Hu, Yi Zhao, Yonghang Tai, and Huafeng Li. [CLS] is not enough: Multi-label recognition via patch-level inference and adaptive aggregation. In *Proceedings of the International Conference on Machine Learning*, 2026.

Ming-Kun Xie, Zhiqiang Kou, Zhongnian Li, Gang Niu, and Masashi Sugiyama. Unlocking the power of co-occurrence in CLIP: A DualPrompt-driven method for training-free zero-shot multi-label classification. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=QGXVZ0OPLy.

Tengyu Yin, Hongmei Chen, Jihong Wan, Pengfei Zhang, Shi-Jinn Horng, and Tianrui Li. Exploiting feature multi-correlations for multilabel feature selection in robust multi-neighborhood fuzzy β covering space. *Inf. Fusion*, 104:102150, 2024.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.*, 26(8):1819–1837, 2014.

Min-Ling Zhang, Qian-Wen Zhang, Jun-Peng Fang, Yu-Kun Li, and Xin Geng. Leveraging implicit relative labeling-importance information for effective multi-label learning. *IEEE Trans. Knowl. Data Eng.*, 33(5):2057–2070, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022.

Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 2027–2036, 2017.

A NOTATION

Table 3 summarizes the notation introduced in Sections 3 and 4. Related quantities are grouped together; dummy indices are defined at first use and omitted here.

Table 3: Notation used in GRADE. Related symbols are grouped by function.

Symbol	Description
$I, \mathcal{C}, C, \mathbf{y}$	Input image, target vocabulary, its size, and evaluation-only binary label vector.
$\mathbf{s}(I), s_c(I)$	Final score vector and its component for class c.
$E_{\mathrm{v}}, E_{\mathrm{t}}, \alpha$	Frozen CLIP visual/text encoders and pretrained logit scale (100 in cached runs).
$M, p_m(c), \mathbf{t}_c$	Number of prompt templates, instantiated prompt, and normalized class prototype.
$\mathrm{norm}, \mathrm{LN}_{\mathrm{post}}, \mathbf{W}_{\mathrm{p}}, \epsilon$	ℓ_2 normalization, post-Transformer normalization, visual projection, and stability constant (10^{-12}).
$\mathbf{g}, a_c^G, G_c(I)$	Normalized global feature, class logit, and vocabulary-normalized global score.
$N, \mathbf{z}_i^{\mathrm{loc}}, \mathbf{u}_i, P_{i,c}$	Patch count, local token, projected normalized patch feature, and patch–class response.
$\mathbf{H}_c, h_c^{\min}, h_c^{\max}, \widehat{\mathbf{H}}_c$	Raw class map, its extrema, and min–max normalized map.
$\mathcal{C}_I, P_{\min}, P_{\max}, \tau_{\mathrm{cand}}, \eta$	Candidate classes, response extrema, candidate coefficient, and map threshold.
$\mathcal{B}_{c,r}, R_c, R$	Retained region box, retained count for class c, and per-class region budget.
$J, \rho_j, I_{c,r}^{(j)}, \mathbf{v}_{c,r}^{(j)}, \ell_{c,r}^{(j)}$	Granularity count, context margin, crop, normalized crop feature, and crop response.
$\ell_{c,r}^{\min}, \ell_{c,r}^{\mathrm{geo}}, \ell_{c,r}^{\max}$	Minimum, geometric-mean, and maximum responses across granularities.
$\mathcal{D}_{\mathrm{ref}}, I_n^{\mathrm{ref}}, N_{\mathrm{ref}}, k_c(I), \xi_c(I)$	Reference subset, its images and size, ECDF count, and normalized global-score rank.
$\mathrm{clip}_{[0,1]}, \mathbb{I}, \widetilde{\ell}_{c,r}, L_c, r_c^\star$	Clipping and indicator operators, calibrated regional response, strongest local score, and its region index.
$D_c, \tau_d, K, \mathcal{N}_K(c)$	Local-evidence fraction, verification threshold, neighbor count, and semantic-neighbor set.
$\mathbf{e}_{c\succ k}, \mathbf{e}_{k\succ c}, \mathbf{d}_{c,k}, q_{c,k}, W_c, w_c$	Comparative prompt embeddings, semantic direction, comparison margin, support fraction, and reliability weight.

B DATASET DETAILS

This appendix records the prepared splits and filtering used for the reported experiments. The counts refer to the files consumed by the evaluation code, not to the nominal size of each source dataset.

MS-COCO. We use the 2014 multi-label split with 80 object categories. The prepared training pool contains 82,081 images and the validation split contains 40,137 images. All methods use the same class order and validation-image order.

VG-256. VG-256 is a 256-category multi-label subset of Visual Genome. The prepared split contains 74,691 training images and 32,011 evaluation images, with a fixed class and image order shared by the aligned methods.

C IMPLEMENTATION DETAILS

Table 4 summarizes the saved GRADE configuration. Both CLIP encoders remain frozen. Global and crop responses are softmax scores over the same target vocabulary; their final sum is not renormalized. Reference images are sampled without replacement using `numpy.random.default_rng(1)`, retaining $\lfloor 0.01 N_{\mathrm{train}} \rfloor$ images and recording their indices.

Here N_{train} is the number of rows in the prepared training annotation file. Sampling follows the annotation-file order; the selected indices are sorted before they are used. The exact 80 prompt strings are the `PROMPT_TEMPLATES` constant in the project implementation and are exported in the accompanying reproducibility artifact; no additional prompt templates are introduced. Their labels are not used by GRADE. All cumulative ablations reuse identical global and crop scores, with the external fusion coefficient fixed to one; SCV additionally applies the class-specific weight w_c. DualCoOp uses the same sampled indices with labels, 50 epochs, batch size 16, and class-agnostic prompts; its saved learning rates are 0.002 for COCO and 0.001 for VG-256.

Crop geometry and numerical stability. Cropping operates on the preprocessed image tensor. Patch-grid boxes are mapped to tensor coordinates using the grid-cell dimensions. For a box of width w and height h, margin ρ extends its left/right boundaries by ρw and its top/bottom boundaries by ρh. Lower coordinates are rounded down and upper coordinates rounded up, then clipped to the tensor bounds. The shorter side is padded to form a square, splitting padding equally except that any extra pixel goes to the right or bottom. Padding uses zero in tensor space, not necessarily black in RGB space. Bilinear interpolation with `align_corners=False` returns each crop to 224×224. Normalization divides by $\max(\|\mathbf{z}\|_2, 10^{-12})$. A zero comparative-text difference remains zero, yielding $q_{c,k} = 0$ and therefore no positive support under the strict comparison in Eq. (15).

Inference cost. Training-free refers to the absence of target-task parameter optimization, not to inference cost. The COCO cache contains 863,114 retained class-specific regions over 40,137 evaluation images. With three context granularities, these correspond to 2,589,342 regional crop evaluations, or 64.51 per image on average, in addition to global and dense-feature extraction. Counts are per retained class–region–granularity tuple, not unique crop coordinates. They exclude reference-image preprocessing and additional semantic-verification work. End-to-end latency and memory consumption have not been benchmarked under a common hardware and batching protocol, so crop counts are not speed measurements.

Evaluation and comparison protocols. mAP is the unweighted mean of class-wise AP from continuous scores. AP uses the project's score-grouped precision–recall implementation with a stable descending score order; no decision threshold is involved. For Macro-F1, we fix a 2% predicted-positive operating point for every reported method. Each method's threshold for each class is the 98th percentile of that class's scores over the entire evaluation split, computed with NumPy's linear quantile interpolation. Scores greater than or equal to the threshold are predicted positive; class-wise F1 scores are then averaged without weighting. The same percentile rule is applied separately to each method and class, yielding approximately 2% predicted positives per class; ties at the threshold can increase this fraction. Ground-truth labels are not used to compute these quantile thresholds, but are used to calculate F1. This protocol accesses the evaluation-score distribution and is not equivalent to fixing thresholds on an independent reference set before deployment. The common prediction fraction need not match each class's true prevalence, so Macro-F1 should be interpreted at this operating point. mAP is computed from continuous scores and is unaffected by this thresholding. P@k averages the fraction of positives among each image's top-k labels.

The prepared annotation lists are consumed in their recorded order. The loader verifies that every listed image exists and raises an error otherwise; no additional image repair or category remapping is applied. Class names are read from the dataset-specific files `coco_labels.txt` and `vg256_labels.txt`. The training and evaluation lists are disjoint by relative image path.

The project adapters retain each upstream method's scoring computation while changing data plumbing, target-vocabulary handling, and score export. For SPARC, we form pair prompts from each class's eight nearest CLIP text neighbors and apply the released debiasing and PCA fusion jointly to all images in each aligned evaluation split; each pair contributes to both named classes. For the 1%-labeled condition, we rank partners by empirical class-conditional co-occurrence in the fixed seed-1 1% training subset and use CLIP text neighbors to fill classes with fewer than eight observed partners. Pair prompts use the template `{c_i} and {c_j}`. Score normalization and PCA still use all aligned evaluation images without their labels. TagCLIP uses the project flat lists and appends its generic COCO background vocabulary after removing duplicate target names. PIAA and PIAA_SC use the same sampled unlabeled-image budget as GRADE; PIAA_SC additionally uses the upstream SC-CLIP patch frontend. DualPrompt derives co-occurrence prompts from the

sampled labeled rows with its published prompt construction and uses the project class order. VG-256 PIAA results are project adaptations; their fixed settings are recorded in the run manifests. All continuous scores are exported before any threshold-dependent metric. DualCoOp selects the checkpoint with highest mAP on the reporting split, not a held-out tuning split. DualCoOp-TF uses no target images, labels, or gradient updates. TagCLIP and PIAA/PIAA_SC on VG-256 are project adaptations. DualPrompt on VG-256 uses 1% rather than its published 2% training fraction. PIAA and PIAA_SC estimate adapters from 1% unlabeled target images, while DualPrompt uses labels to estimate relationships without optimization.

Table 4: GRADE implementation settings. Reference counts are ordered as COCO and VG-256.

Setting	Value
Backbone / encoder input	OpenAI CLIP ViT-B/16 / 224×224
Patch grid / text templates	14×14 / 80-template ensemble
Pretrained similarity scale α (frozen)	100
Candidate threshold τ_{cand}	0.5
Normalized response-map threshold η	0.5
Region extraction	4-connected components, descending component area
Maximum regions per candidate R	5
Context margins $\{\rho_j\}_{j=1}^{J}$	$\{0.1, 0.2, 0.4\}$; $J = 3$
Crop preparation	Preprocessed tensor crops, zero square padding, bilinear resize
Numerical stability ϵ	10^{-12}
Aggregation without DGC	Geometric mean across margins, maximum across regions
Unlabeled reference fraction / seed	1% / 1
Unlabeled reference counts	820 / 746
DGC anchors	Minimum, geometric mean, maximum at $\xi_c = 0, 0.5, 1$
SCV neighbors K / activation threshold τ_d	5 / 0.90
SCV prompt pair	`a photo of {c} rather than {k}` and its reverse
SCV weight when active	$w_c = \min(1, 2W_c)$; otherwise $w_c = 1$
Final fusion / inference seed	$G_c + w_c L_c$ / 0
Macro-F1 operating point	Fixed 2% rate; 98th percentile; linear; $\geq$

D Supplementary Experimental Analysis

D.1 Sensitivity to Local Evidence Budgets

We reaggregate saved crop responses on all 40,137 COCO validation images, without rerunning CLIP. At $R = 5$, we evaluate every nonempty subset of $\{0.1, 0.2, 0.4\}$; we also vary R from one to five for $J = 1$ and $J = 3$. Regions retain their original descending-area order. DGC and SCV are disabled throughout, and scores are fused as $G + L$, isolating local aggregation from calibration. The geometric mean uses the selected J margins. The three configurations shared with Table 2 reproduce its mAP values.

Context diversity. At $R = 5$, three margins improve mAP by 0.23 percentage points over the best single margin, but by only 0.015 points over the best pair $\{0.2, 0.4\}$. Moreover, $\{0.1, 0.2\}$ slightly underperforms the single margin 0.4: adding crops alone does not ensure better ranking. These comparisons are consistent with complementary context being more useful than the number of views itself. The third margin adds 50% more crop evaluations than a pair for a small mAP gain, supporting $J = 3$ as a coverage-oriented default rather than a uniquely optimal choice.

Regional coverage. With three margins, increasing R from one to five gains 0.79 mAP points; $R = 3$ already captures 88.2% of this improvement. Increasing the budget from three to five adds only 0.093 points while increasing retained crop evaluations by 27.4% (2,033,217 at $R = 3$ versus 2,589,342 at $R = 5$, counted over the full COCO evaluation set without deduplicating coordinates). This count increase is not an end-to-end latency increase. The similar flattening under a single margin suggests that diminishing returns are not specific to multi-granularity fusion. Thus, $R = 5$ favors coverage, whereas $R = 3$ offers a lower-cost alternative. These are descriptive comparisons within the saved margins and $R \leq 5$; they neither establish statistical significance nor test larger budgets.

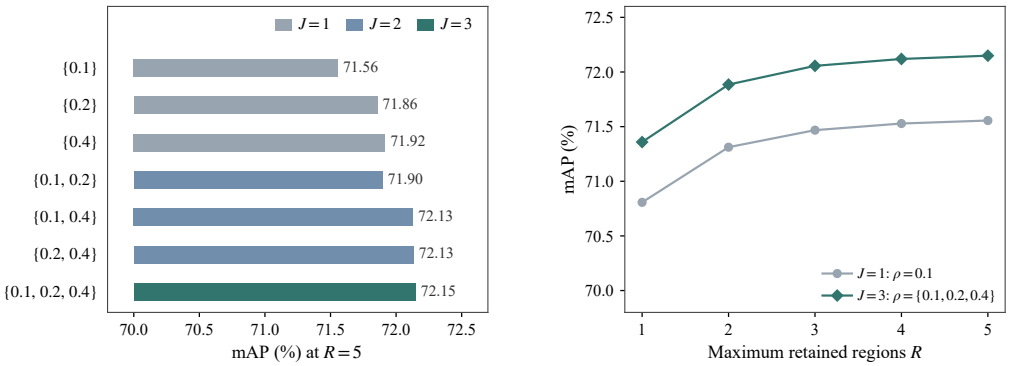


Figure 6: COCO sensitivity to context margins (left, $R = 5$) and region budget (right). Each left-hand tick lists the margins used together. All settings use uncalibrated local aggregation and unit fusion weight. The left mAP axis is truncated; bars start at 70%.

D.2 Conditional Calibration Effects

Figure 5 resolves the aggregate changes at class level. The regional stage produces broad positive shifts on COCO and VG-256 (more than 90% of classes improve). DGC remains centered near zero but is positive for 70–78% of classes, consistent with targeted distribution correction rather than a uniform gain. SCV is even more concentrated near zero and slightly negative on VG-256, showing that semantic calibration trades gains and losses across classes after coverage has been established.

DGC adds 0.408 and 0.322 pp mAP on COCO and VG-256, respectively, but reduces COCO P@3 by 0.14 pp. Class-wise ordering across images and label ordering within an image therefore need not improve together. SCV provides a further 0.063 pp mAP on COCO and costs 0.052 pp on VG-256 despite slightly higher P@1 and P@3. This limits the interpretation of the qualitative examples: selective suppression can resolve individual confusions without yielding a uniform aggregate gain. Semantic preference is a fallible reliability proxy, not a correctness test. These cumulative comparisons measure gains conditional on preceding stages; they establish neither independent component effects nor statistical significance.

D.3 COCO Aggregation and Verification Diagnostics

We further examine DGC and SCV on the full COCO evaluation split using saved predictions, without additional model inference. Labels are used only for this retrospective evaluation.

Fixed aggregation controls. We retain the same candidates, three context scales, and up to five regions. Only aggregation across scales within each region changes. All variants then take the maximum across regions and use $G + L$ with SCV disabled. DGC uses the existing 1% unlabeled reference subset, whereas fixed aggregation requires no reference images.

Table 5: Aggregation controls on COCO (%). All variants reuse the same regional predictions.

Aggregation	mAP	P@1	P@3	P@5
Minimum	71.61	91.46	52.96	37.51
Arithmetic mean	72.34	**91.64**	**53.23**	37.77
Maximum	72.29	91.51	52.84	37.67
Geometric mean	72.15	91.56	53.20	37.72
DGC	**72.56**	91.61	53.06	**37.78**

DGC improves mAP by 0.219 pp over the strongest fixed control, arithmetic mean, with higher AP for 51 of 80 classes. Its gain over geometric mean is 0.408 pp. However, arithmetic mean yields

higher P@1 and P@3, so the mAP advantage does not imply improvement under every ranking metric.

SCV selectivity and positive-label attenuation. We compare image–class candidates satisfying the pre-verification condition $D_c \geq 0.90$. This fixes the eligible population before inspecting SCV's effect. A candidate is attenuated when $w_c < 1$. Positive and negative refer to the dataset labels, not independently verified object presence or absence.

Table 6: SCV effects among eligible COCO candidates. Rates use the count in each label group as the denominator.

Label group	Eligible	Attenuated	Rate (%)	Mean w_c
Positive	30,138	1,749	5.80	0.977
Negative	169,234	47,127	27.85	0.864

Negative-label candidates are attenuated more often than positive-label candidates. Averaging rates equally across the 80 classes gives 28.62% versus 7.45%, respectively, and the negative-label rate is higher in 79 classes. These observations support selective attenuation on COCO, while the 1,749 attenuated positive-label candidates demonstrate imperfect selectivity. Lowering a negative-label score does not necessarily correct a ranking error, and incomplete annotations can affect this diagnostic. Consistent with this distinction, SCV adds only 0.063 pp mAP on COCO and reduces VG-256 mAP by 0.052 pp in Table 2. These descriptive analyses establish neither statistical significance nor generalization beyond the evaluated setting.

E AI USE STATEMENT

Generative AI tools were used to assist with a limited range of supporting tasks during the research process, including checking research code, reviewing dataset processing and evaluation procedures, assisting with the analysis and interpretation of experimental results, and improving the language and presentation of the manuscript. The research methodology and main experimental design were developed by the authors. AI tools also assisted with executing supplementary analyses of saved prediction caches. All numerical results reported in this manuscript were obtained from the actual experiments described in the paper and were not generated or inferred by generative AI. The authors reviewed and verified all AI-assisted code, analyses, and textual content and take full responsibility for the research content, experimental results, and final submitted version of the manuscript.