WHEN DO LOCAL FEATURES HELP? PATCH-AREA MATCHING IN FINE-GRAINED VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive vision–language models (VLMs) align a single global image embedding with text, capturing whole-image structure but discarding local detail that fine-grained recognition depends on. Frozen patch features retain localizable structure that text-based scoring misses (Zhou et al., 2022a; Wang et al., 2024a), while gradient-based adaptation risks eroding transferable features (Kumar et al., 2022; Wortsman et al., 2022). We propose **GL-VLM**, a *training-free, few-shot* method that keeps the backbone frozen and augments the global embedding with a local *patch-area* descriptor: per class, we cluster the support patches into region prototypes and score a query by how its patches fall into each class's areas, combined with a global prototype. Stacked on a strong Tip-Adapter, it improves texture/part/scene categories (up to $+5.3$ on DTD at k=16, and a texture/scene mean of $+8.7$ over 15 backbones at k=5; significant on the texture datasets at $p<0.05$), with a β=0 validation-safe fallback that empirically matches or improves the baseline under our validation-selected protocol. The benefit is interpretable and stated a priori: local patch-areas help where distinctions are textural, part-based, or scene-level, and are neutral where text is near-perfect or the cue is global shape. Text priors and local structure are complementary, and which wins is predictable from the category. Because the backbone is never updated, its pretrained zero-shot representation is left unchanged, at zero training cost. We do not claim state of the art (covariance-based methods match or exceed it), but characterize *when* local features help.

1 INTRODUCTION

A contrastive VLM such as CLIP (Radford et al., 2021) aligns a single pooled image embedding $\boldsymbol{g}$ with text, capturing whole-image and shape structure but discarding the local, textural, and part-level detail that separates fine-grained categories. Frozen patch features retain this discriminative signal even though text-based (zero-shot) scoring does not surface it (Zhou et al., 2022a; Wang et al., 2024a), but recovering it appears to require labels rather than text matching. This raises the question we take as our title: *when* do local features help fine-grained recognition, and can we predict where?

Two facts constrain the answer. First, fine-tuning the encoder can distort pretrained representations and degrade out-of-distribution transfer (Kumar et al., 2022; Wortsman et al., 2022; Ming & Li, 2024). Second, collapsing patches into one vector does not help; the signal is distributed across patches. We therefore keep the backbone *frozen* and add the local signal at the patch level, using a few labels.

GL-VLM (patch-area matching). For each class we cluster the patch features of its few support images into M region prototypes (the *areas* the class occupies in the shared space), and score a query by how well its patches fall into each class's areas (a late-interaction match), combined with the global few-shot prototype. The method is training-free in the sense of requiring no gradient-based optimization (the backbone is frozen and only k-means runs at test time, though β is selected from labeled validation data), and strictly generalizes the global classifier.

Contributions.

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.

- A training-free, few-shot local descriptor for frozen VLMs based on per-class patch-region prototypes and late-interaction scoring, which *stacks on a strong training-free few-shot baseline*: adding it to Tip-Adapter improves texture/scene categories (up to $+5.3$ on DTD at $k{=}16$; significant at $p{<}0.05$) and, under a $\beta{=}0$ validation-safe fallback, matches or improves it, reaching accuracy competitive with a *trained* adapter (Tip-Adapter-F) on texture/scene (Sec. 5, Table 1; per-backbone results in the supplementary).
- A predictable *complementarity* between the text prior and local structure: text-nameable categories and low-shot favor text, while texture/scene/fine categories at $k{\geq}5$ favor local patch-areas, beating even Tip-Adapter (Table 2).
- An interpretable finding, with cue-type predictions stated before measurement, isolating the local contribution: patch-areas help texture/part categories and are neutral for global-shape categories, across the full dataset suite (Table 3).
- A deployment property: because the backbone is never updated, the foundation model and its pretrained zero-shot representation are left *unchanged* (no catastrophic forgetting), unlike gradient-based adaptation such as prompt tuning or full fine-tuning. This is shared with other training-free methods and comes here at zero training cost.
- An analysis showing the same patch-area classifier improves *open-set* recognition on texture/scene categories across 15 backbones, though text remains the stronger overall rejector (Sec. 8).
- An explicit scope statement: the method is label-conditioned and does not address zero-shot, retrieval, or generative tasks (Sec. 4).

2 Related Work

Training-free few-shot CLIP. Tip-Adapter (Zhang et al., 2022) adapts CLIP to few-shot classification without gradient training by building a key–value cache from support features; prototype- and cache-based variants follow the same global-feature recipe, as does prototypical few-shot learning (Snell et al., 2017). Our method differs by operating on *dense patch* features clustered into per-class region prototypes, rather than on the global embedding, and combines a late-interaction patch-to-region score with the global prototype. It is thus a strictly larger hypothesis class that recovers the global cache method at $\beta{=}0$.

Local and patch-level alignment. Patch- and region-text alignment during pretraining (Yao et al., 2022; Zhong et al., 2022; Bica et al., 2024; Li et al., 2022b) and prototype/clustering objectives (Li et al., 2021; Caron et al., 2020; Wang et al., 2022) recover spatial detail, but require training and modify the pretraining objective. FILIP's token-wise maximum similarity (Yao et al., 2022) is the late-interaction mechanism we adapt to the few-shot region-matching setting at test time; late interaction originates in retrieval (ColBERT (Khattab & Zaharia, 2020)) and has been applied to document VLMs (ColPali (Faysse et al., 2025)). Region-token methods on frozen backbones (Khosla et al., 2026) pursue related goals *with* training; we remain training-free. Recent training-free few-shot methods refine the global class prototype, for example via a shared-covariance discriminant analysis (Wang et al., 2024b) or cross-modal prototype alignment and mixing (Goswami et al., 2026); we instead add a per-class *patch-region* score, which stacks on such stronger prototypes (Table 1). Our cue-type axis connects to the documented shape-vs-texture bias of encoders (Gavrikov et al., 2025).

Dense features from frozen CLIP. Training-free methods such as MaskCLIP and SCLIP (Zhou et al., 2022a; Wang et al., 2024a) show that pretrained CLIP patch features retain localizable signal without gradient-based adaptation; ProxyCLIP (Lan et al., 2024) reaches similar ends by borrowing spatial correspondence from other vision foundation models. Related *training-based* approaches instead exploit patch/region structure through learned objectives (Xu et al., 2022; Luo et al., 2023; Bica et al., 2024). We use MaskCLIP-style dense features (Zhou et al., 2022a) as our patch view.

Robust adaptation. Because fine-tuning can distort pretrained features and erode zero-shot transfer (Kumar et al., 2022; Wortsman et al., 2022; Ming & Li, 2024), we avoid training the backbone entirely and add local evidence post hoc, at the representation level.

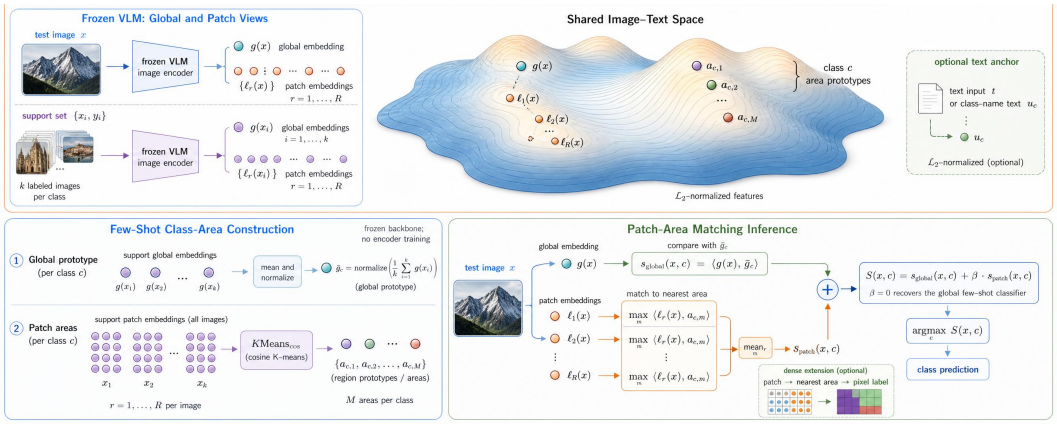

Figure 1: **GL-VLM overview (training-free).** *Left:* a frozen VLM image encoder maps the test image and the k labeled support images per class to a global embedding $\boldsymbol{g}$ and dense patch embeddings $\{\boldsymbol{\ell}_r\}$, all L_2-normalized in the shared image–text space; an optional class-name text anchor $\boldsymbol{u}_c$ can replace the global prototype. *Bottom-left:* per class c we form a global prototype $\bar{\boldsymbol{g}}_c$ (mean of support globals) and M *area* prototypes $\{\boldsymbol{a}_{c,m}\}$ by cosine k-means over the support patches; the backbone is never trained. *Right:* a query is scored by a global term $s_{\text{global}}(x, c) = \langle \boldsymbol{g}(x), \bar{\boldsymbol{g}}_c \rangle$ plus a late-interaction patch-to-area term $s_{\text{patch}}(x, c) = \text{mean}_r \max_m \langle \boldsymbol{\ell}_r(x), \boldsymbol{a}_{c,m} \rangle$, combined as $S(x, c) = s_{\text{global}} + \beta\, s_{\text{patch}}$ (with $\beta = 0$ recovering the global few-shot classifier). The per-class areas can also be applied densely (patch $\rightarrow$ nearest area $\rightarrow$ pixel label), which we leave to future work.

3 Method

For a frozen VLM and image x: the **global view** $\boldsymbol{g}(x) \in \mathbb{R}^d$ is the standard image embedding; the **patch view** $\{\boldsymbol{\ell}_r(x)\}_{r=1}^{R}$ are MaskCLIP-style dense patch features (Zhou et al., 2022a), projected to the shared space and L_2-normalized.

Learning class areas (few-shot). For each class c, gather the patches of its k support images and cluster them into M region prototypes, the areas class c occupies:

$$\{\boldsymbol{a}_{c,1}, \ldots, \boldsymbol{a}_{c,M}\} = \text{KMeans}_{\cos}\big(\{\boldsymbol{\ell}_r : \text{image} \in \text{support}(c)\}\big). \tag{1}$$

Also form the global prototype $\bar{\boldsymbol{g}}_c = \text{normalize}(\text{mean}_{x \in \text{support}(c)} \boldsymbol{g}(x))$.

Scoring.

$$s_{\text{global}}(x, c) = \langle \boldsymbol{g}(x), \bar{\boldsymbol{g}}_c \rangle, \tag{2}$$

$$s_{\text{patch}}(x, c) = \frac{1}{R} \sum_{r=1}^{R} \max_m \langle \boldsymbol{\ell}_r(x), \boldsymbol{a}_{c,m} \rangle, \tag{3}$$

$$S(x, c) = s_{\text{global}}(x, c) + \beta\, s_{\text{patch}}(x, c). \tag{4}$$

The $\max_m$ selects the best-matching region per patch (local evidence); the mean_r aggregates over patches. Setting $\beta = 0$ recovers the global few-shot classifier, so (4) strictly generalizes it. The method is training-free: only k-means on support patches is computed at test time.

Design rationale. The backbone is frozen because trained variants lost to frozen features in prior work. Labels are required because the fine structure is reachable only with labels, not by text scoring. On the number of areas: a single per-class centroid ($M = 1$) already recovers much of the gain (Table 4), and multiple areas add a smaller further increment on the most texture-heavy datasets. The essential operation is therefore *not* multimodal region discovery per se, but retaining patch-level query structure and applying late interaction against class-conditioned patch prototypes; we do *not* claim that per-region structure ($M > 1$) is the sole source of the gain. Concretely, matching patches to a small set of class centroids (rather than pooling patches into one vector, which collapses to the $\beta = 0$ baseline; Table 4) is what makes the local term usable. $M = 8$ is a robust default (plateau for $M \geq 8$).

4 APPLICABILITY ACROSS DOWNSTREAM TASKS

GL-VLM is *label-conditioned*: it builds per-class areas from labeled support patches, which determines its scope. The core use is few-shot fine-grained classification, where the areas serve directly as a classifier; it extends partially to weakly-supervised localization (patch→area gives a coarse relevance map) and, as future work, to dense segmentation. Because the areas are per-class and constructed from labels, the method does not apply to zero-shot classification (no labels to build areas from; the global term at β=0 is the underlying few-shot baseline it generalizes, not zero-shot CLIP), image–text retrieval (which uses the global embedding and has no per-class formulation), or generative tasks. We report GL-VLM as complementary to strong frozen few-shot baselines, with the where-it-helps analysis as the central contribution.

5 EXPERIMENTS

Frozen CLIP ViT-L/14, M=8 areas, few-shot episodes averaged over standard splits; we report $k \in \{1, 5, 10, 16\}$ (Table 2 gives the shot sweep). We report gains under two protocols and label which applies: the ***CoOp protocol*** (official few-shot splits, 50 episodes, used for Tables 1–2) and an *episode protocol* (random N-way episodes over 15 backbones, used for the per-backbone generalization results); k=16 numbers (e.g. DTD +5.3 on CLIP ViT-L/14) are always labelled as such. **The fusion weight β is selected on a held-out validation split disjoint from both the support and the test set.** For each episode, the k support images per class are used to build the classifier, while a *disjoint set of labeled images from the same classes* is used only to select β by validation accuracy; the fixed β is then evaluated on the test set, which is never used for hyperparameter selection (the validation split holds 16 images per class, drawn from the same support classes). Because β=0 recovers the base classifier and β is never tuned on test data, the β=0 fallback is *validation-safe*: the selected configuration is never worse than the baseline *on the validation split*. This is an empirical property of validation selection, not a test-set guarantee (validation selection can overfit), so we report the test-set outcome separately. Cue-type labels are assigned *a priori* from category definitions before observing gains, so the texture-vs-shape prediction is a genuine test rather than a post-hoc line. To reduce subjectivity, we also verify the labels against a measurable proxy: a category is "nameable" when zero-shot text scoring already classifies it well (high zero-shot accuracy: Food-101, Pets), and "texture/scene" when zero-shot text is weak but image geometry is strong (low zero-shot accuracy, high linear-probe accuracy: DTD, RESISC45); the a-priori labels agree with this proxy on all seven datasets. This axis aligns with the documented shape-vs-texture bias of vision encoders (Gavrikov et al., 2025). Combined-score means with episode standard deviations (across 100 episodes, k=5) are DTD 0.764 ± 0.025, Flowers 0.981 ± 0.007, and RESISC45 0.886 ± 0.020 (reported as mean $\pm$ std, not confidence intervals); the texture/scene gains, though modest, are consistent across episodes and backbones.

Reproducibility and the choice of β. Patch features are the pre-projection dense tokens (L_2-normalized); the token count R is the backbone grid (196/256/576 for ViT-B/16 / L/14 / 336px; 32 Q-Former queries for BLIP-2), tabulated in the supplementary. Areas use cosine k-means (k-means++, M=8, 10 restarts). Our default selects β on a held-out validation split; because that uses labels, we also report a strict K-shot protocol in which β is selected by leave-one-shot-out cross-validation within the K shots (no extra labels), or fixed to a single value; in all cases the β=0 fallback keeps the configuration validation-safe (never worse than the baseline on the selection split).

5.1 MAIN RESULT: PATCH-AREAS IMPROVE A STRONG TRAINING-FREE BASELINE

Our primary method adds the patch-area score on top of Tip-Adapter (Zhang et al., 2022), a strong and widely used training-free few-shot baseline:

$$S(x, c) = \mathrm{TipAdapter}(x, c) + \beta\, s_{\text{patch}}(x, c), \tag{5}$$

so β=0 recovers Tip-Adapter exactly and the method can only add, never subtract, in expectation over β selection. The per-backbone episode-protocol gains (15 backbones, CLIP ViT-L/14) are reported in the supplementary and agree with the standard-benchmark result below.

Table 1: Few-shot accuracy (%) at k=16, CoOp protocol (ViT-B/16, full test sets, 3 seeds). GL-VLM improves the training-free Tip-Adapter on texture/scene and is neutral on shape/nameable, but covariance methods (GDA, TAMP+LDA) and the trained Tip-Adapter-F match or exceed it.

Method	DTD	Flow.	Cars	Airc.	Pets	Food
Tip-Adapter	62.7	91.0	73.3	37.3	89.8	88.6
Tip-X	65.5	93.3	76.1	39.8	91.8	89.2
APE	67.8	92.7	76.4	39.3	90.2	88.5
TAMP+LDA	70.0	96.9	81.2	46.5	92.5	89.0
GDA	71.7	97.4	81.8	48.5	91.7	87.7
Tip-Adapter-F (trained)	72.4	95.9	81.9	47.1	92.7	89.4
GL-VLM (ours)	70.9	94.8	74.3	37.7	90.0	89.0

Headline. A single training-free rule improves a strong training-free few-shot baseline: the gain (up to $+5.3$ on DTD at k=16, CoOp protocol; and a texture/scene mean of $+8.7$ over 15 backbones under the episode protocol at k=5) falls on texture/scene/fine categories where text is a poor descriptor, and is within episode noise where text is already near-perfect (Food, Pets), with a β=0 validation-safe fallback. Consistent, significant on the texture datasets ($p<0.05$), and predictable from cue-type, it matches a *trained* adapter on texture/scene without any training (trailing it on shape datasets).

5.2 Text and local structure are complementary

The pattern reflects a *predictable complementarity*: text-nameable categories (Pets, Food) and low-shot (k=1) favor the text prior, while texture/scene/fine categories (DTD, Flowers, RESISC45) at $k\geq5$ favor local patch-areas. Local evidence helps where the text prior is a weak descriptor and is neutral on global shape and text-nameable categories; Table 3 compares the label-free combination (global + patch-areas, no text prior) against Tip-Adapter directly.

5.3 Comparison on the standard few-shot benchmark

We evaluate on the CoOp protocol (Zhou et al., 2022b) (ViT-B/16, official few-shot splits, full test sets, mean over 3 seeds) against training-free few-shot methods (Tip-Adapter (Zhang et al., 2022), Tip-X (Udandarao et al., 2023), APE (Zhu et al., 2023), GDA (Wang et al., 2024b), TAMP+LDA (Goswami et al., 2026)) and the trained Tip-Adapter-F. GL-VLM is our patch-area term stacked on Tip-Adapter, with the fusion weight β selected on a held-out split (falling back to β=0 when it cannot be validated, e.g. at k=1).

We report the comparison honestly. GL-VLM improves the training-free Tip-Adapter on texture/scene categories (Table 2) but does *not* set a new state of the art: covariance-based training-free methods (GDA, TAMP+LDA) and the trained Tip-Adapter-F match or exceed it at standard shots, most clearly on shape datasets (Cars, Aircraft) where the patch term is inactive. The value of GL-VLM is not peak accuracy but the interpretable, predictable *where*-it-helps pattern and what it reveals about the local signal (Sec. 6).

The patch term is a baseline-independent booster. Stacking the patch-area score on *each* of the five bases (Tip-Adapter, GDA, Tip-X, APE, TAMP+LDA) reproduces the same dissociation: at k=16 on DTD it lifts every base ($+8.0$ tip, $+6.4$ Tip-X, $+4.3$ APE, $+3.2$ TAMP, $+1.6$ GDA; Flowers similar) and is ≈ 0 on shape/nameable for all five (full 5×7 table in the supplementary). But best-base + patch beats the best base alone by only $+1.7$ (DTD) and $+0.2$ (Flowers) at k=16: GDA and TAMP+LDA already start high, so the patch signal, though real and general, adds most to weaker bases and cannot lift the strongest past their ceiling.

Table 2: Gain of GL-VLM (Tip+patch) over Tip-Adapter (%), CoOp protocol, ViT-B/16, mean over 50 episodes (5 splits $\times$ 10 seeds). The patch term helps texture/scene categories, growing with shots, and is ≈ 0 on shape and nameable categories. Boldface marks gains that are significant at $p<0.05$ (paired t-test over episodes); the two texture datasets are significant at every shot, while the shape/nameable gains are within noise. This is the cue-type dissociation.

Dataset	cue-type	k=1	k=2	k=4	k=8	k=16
DTD (Cimpoi et al., 2014)	texture	$\mathbf{+0.5}$	$\mathbf{+2.3}$	$\mathbf{+2.7}$	$\mathbf{+4.9}$	$\mathbf{+5.3}$
Flowers-102 (Nilsback & Zisserman, 2008)	texture/fine	$\mathbf{+0.6}$	$\mathbf{+1.1}$	$\mathbf{+1.5}$	$\mathbf{+1.1}$	$\mathbf{+0.6}$
RESISC45 (Cheng et al., 2017)	texture/scene	$\mathbf{+1.0}$	$\mathbf{+1.1}$	$\mathbf{+0.9}$	$\mathbf{+1.4}$	$\mathbf{+1.0}$
Stanford-Cars (Krause et al., 2013)	shape	+0.0	+0.0	-0.1	+0.1	+0.6
FGVC-Aircraft (Maji et al., 2013)	shape	+0.0	+0.0	-0.1	-0.0	+0.9
Oxford-Pets (Parkhi et al., 2012)	nameable	-0.1	+0.1	+0.1	+0.3	-0.1
Food-101 (Bossard et al., 2014)	nameable	+0.0	+0.0	-0.1	-0.1	-0.1

Table 3: Ablation (no text prior): per-dataset few-shot gain of global+patch-areas over global-only (k=5). Cue-type is assigned a priori; gain > 0 for texture/part rows, ≈ 0 for global-shape rows.

Dataset	Cue-type (pre-reg.)	global	patch-areas	combined	gain	matches?
DTD (Cimpoi et al., 2014)	texture	0.714	0.742	0.764	+0.050	yes
Flowers-102 (Nilsback & Zisserman, 2008)	texture/fine	0.977	0.961	0.981	+0.004	partial
RESISC45 (Cheng et al., 2017)	texture/scene	0.879	0.856	0.886	+0.007	yes
FGVC-Aircraft (Maji et al., 2013)	shape	0.439	0.281	0.438	−0.001	yes
Stanford-Cars (Krause et al., 2013)	shape	0.815	0.633	0.816	+0.002	yes
Oxford-Pets (Parkhi et al., 2012)	nameable	0.795	0.780	0.824	+0.028	**no**
Food-101 (Bossard et al., 2014)	nameable	0.964	0.908	0.964	−0.000	yes

5.4 The cue-type dissociation

Table 2 isolates the patch contribution: it is positive and significant on texture/scene at every shot (DTD +5.3, RESISC45 +1.0, Flowers +0.6 at k=16, all $p<0.05$), and is within episode noise on shape and nameable categories. The gain over Tip-Adapter appears exactly where local structure carries the class signal, and vanishes elsewhere.

5.5 Ablation: where local features help without a text prior

Table 3 isolates the local contribution using the label-free combination $S = s_{\text{global}} + \beta\, s_{\text{patch}}$ (no text prior), the cleanest test of the a-priori cue-type prediction.

5.6 Ablations

Ablation takeaways. M=8 areas saturates (little change from M=4 to M=16; across $M \in \{8, 12, 16, 24\}$ the texture/scene gain stays within episode noise). Only the mean-max late-interaction aggregation works: in our ablation, mean-pool and max-max select β=0 and therefore recover the baseline (no usable patch signal), confirming that keeping patches as regions and matching them individually is what carries the gain. The patch term selects $\beta>0$ under *both* the prototype and the class-text anchor, so the local contribution is anchor-robust; the class-text anchor raises absolute accuracy on nameable categories (Pets, Food) where zero-shot text is strong.

Shot count, and a closed-set/open-set asymmetry. Sweeping the support shots (CoOp protocol, 50 episodes; full sweep in the supplementary), the closed-set DTD gain *grows* with k (+0.5 at k=1 to +5.3 at k=16 on CLIP ViT-L/14), because more support patches yield better-populated areas, and the texture/scene mean plateaus near +2.3. The open-set classification lift moves the opposite way, *shrinking* from +2.6 at k=5 to +2.0 at k=16 (mean over 15 backbones): as the base classifier strengthens, less headroom remains for the patch term. We report closed-set at k=16 (the standard few-shot count) and open-set at k=5.

Strict K-shot β and computational overhead. The supplementary reports the texture/scene gain when β is chosen without any extra validation labels: by leave-one-shot-out cross-validation within

Table 4: Ablations (k=5, CLIP ViT-L/14). M = number of areas; aggregation; β; global anchor. These use a separate 50-episode sample, so absolute values differ from Table 3 (100 episodes) by up to 0.003; trends are identical. Aircraft selects $\beta^\star$=0 throughout (shape; the patch term is switched off), so its column is constant at the global-only value.

Ablation	Setting	Aircraft	Cars	Pets	DTD
Number of areas M	M=1	0.438	0.814	0.804	0.748
	M=4	0.438	0.814	0.830	0.758
	M=8 (default)	0.438	0.815	0.827	0.767
	M=16	0.438	0.818	0.830	0.766
Aggregation	mean-max (ours)	0.438	0.815	0.827	0.767
	mean-pool	0.438	0.814	0.793	0.718
	max-max	0.438	0.814	0.793	0.718
β	β=0 (baseline)	0.438	0.814	0.793	0.718
	$\beta^\star$ (best)	0.438	0.815	0.827	0.767
Global anchor	prototype $\bar{\boldsymbol{g}}_c$	0.438	0.815	0.827	0.767
	class text $\boldsymbol{u}_c$	0.404	0.854	0.911	0.813

the K shots, and with a single fixed β across all datasets. The gain is preserved and the β=0 fallback remains validation-safe (never worse than the baseline on the selection split). The supplementary reports the wall-clock and memory overhead of building and matching areas relative to Tip-Adapter, as a function of shots and resolution; the cost is dominated by one-time k-means over support patches and is a small fraction of feature extraction.

Dense patch tokens vs. re-encoded crops. At an equal count of local features (64), dense MaskCLIP tokens give a texture/scene gain of $+9.7$ versus $+3.7$ for an 8×8 crop grid, and finer crops *hurt*. The advantage of dense tokens is whole-image *context* (each token attends over the whole image), not spatial resolution, justifying dense features as the patch view.

6 THE A-PRIORI PREDICTION, STATED ONCE

To avoid any ambiguity, we state a single prediction and hold to it throughout. We assign each dataset one cue-type from its category definitions, *before* observing gains, using the same labels in every table: **texture/scene** = DTD, RESISC45; **texture/fine** = Flowers-102; **shape** = Aircraft, Cars; **nameable** = Oxford-Pets, Food-101. The prediction is: the combined score beats the baseline on the texture-type datasets (DTD, RESISC45, Flowers), and is within noise of zero on shape and nameable datasets. We count a prediction as *supported* when the mean gain is positive and larger than one episode standard deviation (our proxy for excluding zero); otherwise we count it as within noise. We do not have a timestamped public pre-registration; "a priori" means only that the labels and prediction were fixed before the corresponding numbers were computed, and we report every outcome including the misses below.

Outcome. Two claims must be kept distinct. First, the independently defined cue-type *labels* agree with the measurable proxy (zero-shot vs. linear-probe accuracy) on all seven datasets. Second, the *performance prediction* (gain above one episode std on texture, within noise on shape/nameable) is supported on five of seven datasets. The two deviations: Oxford-Pets (predicted ≈ 0 as nameable) shows a small positive $+0.028$ without a text anchor, plausibly because breed differences are partly fur-texture; and Flowers-102, predicted texture, is only weakly positive without a text prior ($+0.004$, within one episode std and so *not* supported by our criterion) although clearly positive once a text prior is added. With only seven datasets we treat this as a suggestive capability signature rather than a high-powered test, and report per-dataset numbers (Table 3) so readers can judge directly. Two mechanism checks support the reading: GL-VLM improves on Tip-Adapter rather than matching it (Table 2), and removing the patch term (β=0) removes the gain (Table 4).

7 GENERALIZATION ACROSS VLM FAMILIES

GL-VLM requires only dense patch features and a shared space in which a global prototype and patch features are comparable; it needs no text tower for the core mechanism. It therefore applies not only to contrastive dual-encoders but also to generatively trained VLMs that expose patch-level features through a contrastive/ITC interface: CoCa (Yu et al., 2022), BLIP (Li et al., 2022a), and BLIP-2 (Li et al., 2023). (Decoder-only models that produce text rather than a shared embedding space are out of scope, as in Sec. 4.)

Finding: a small, consistent, cue-type-predictable gain across all families. The patch-area gain is positive on texture/scene for essentially every backbone (texture/scene mean $+8.7$ over 15 backbones at k=5, episode protocol; $+2.4$ at k=16, 3-seed), and ≈ 0 on shape and nameable categories. The effect is not confined to a family or to high-resolution features: it appears on contrastive ViTs, CNNs (ConvNeXt, ResNet), sigmoid (SigLIP), EVA, and generative backbones via their contrastive heads (CoCa, BLIP, BLIP-2). The magnitude varies with how much room the strong α-tuned few-shot baseline leaves rather than with patch rawness; where that baseline is already very strong (e.g. SigLIP2 at 384 resolution) the added gain is smallest. The consistent cue-type dissociation, not the magnitude, is the finding: patch-areas help where local structure carries the class, across both contrastive and generative VLMs.

The full per-backbone numbers are given in the supplementary material.

Protocol. For each backbone we extract its native patch features, build per-class areas by k-means (Sec. 3), and report the Tip+patch gain over Tip-Adapter, grouped by cue-type. All hyperparameters (M, aggregation, β selection) are held fixed across backbones. The texture/scene gain is positive on essentially all 15 backbones and neutral on shape/nameable, confirming the cue-type dissociation generalizes across architecture families.

8 OPEN-SET RECOGNITION: A CATEGORY-SPECIFIC BENEFIT

Open-set recognition (OSR) with frozen VLMs is known to be difficult on fine-grained vocabularies the signal separating known from unknown classes is weakly captured by text-based scoring. This raises the question of whether the patch-area classifier, which improves closed-set fine-grained accuracy, also improves OSR. We answer it honestly: it does, but only on the categories where local structure carries the signal, and *not* as a general-purpose rejector.

Protocol. We follow the standard OSR protocol on all seven datasets: half the classes are held out as unknown at random, averaged over five splits. Known classes provide the few-shot support; unknowns appear only at test time. We report AUOSCR (which credits a method only for inputs it both retains and classifies correctly) and OpenAUC.

Text remains the strongest overall rejector. Figure 2 and the supplementary compare GL-VLM against text (MSP, ZeroShot) and geometry-on-patches (KNN, GHOST (Rabinowitz et al., 2025), Mahalanobis on the same patch features), averaged over datasets. **We do not claim to beat text overall**: MSP is the top method on 14 of 15 backbones (by AUROC), and GL-VLM (Tip+patch) typically ranks second or third, trailing MSP by up to 10.5 AUOSCR (and edging it by 0.6 on one backbone). Geometry-on-patches (KNN/GHOST) and the fused score are weaker still. The honest top line is that strong text scoring (MSP over softmaxed text similarities, $\max_c p(y{=}c \mid x)$) is the strongest rejector among the evaluated methods on these benchmarks.

But the patch classifier improves OSR on texture/scene, everywhere. The aggregate hides a consistent per-category effect. Holding the rejection score fixed, replacing the Tip-Adapter classifier with Tip+patch improves AUOSCR on the texture/scene categories by $+2.6$ AUOSCR (mean over 15 backbones; 77.2 vs. 74.6) and is ≈ 0 on shape/nameable (Table 2). This is the *same capability signature* as the closed-set result: patch-areas help where local structure carries the class, now in the open-set classification rate. Full per-method rankings and per-backbone lifts are in the supplementary.

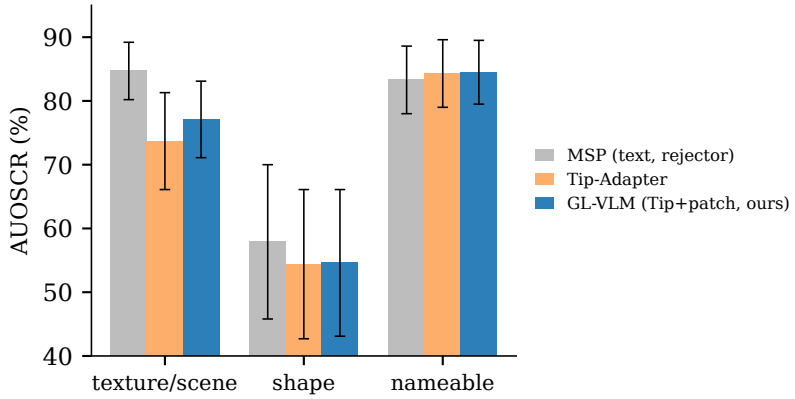

Figure 2: Open-set AUOSCR by cue type (mean over 15 backbones, k=5; error bars are std across backbones). On texture/scene, GL-VLM (Tip+patch) lifts the open-set classification rate over Tip-Adapter by $+2.6$, but text scoring (MSP) remains the stronger *rejector*; on shape and nameable all methods are comparable. Patch-areas improve classification, not rejection. Numeric values in the supplementary material.

What we claim, and an a-priori decisive test. We claim a category-specific improvement to the *classification* component of OSR, not a better rejector than text. We ran a make-or-break test with three controls committed in advance: report AUROC (pure rejection) alongside AUOSCR; decouple classifier from rejector via a shared global+patch head; and tune the fused text/coverage score *only on a held-out pseudo-unknown split* (γ=0 recovers MSP), with the rule that fusion must beat MSP on texture/scene. It does not: fusion beats MSP on $3/12$ texture/scene cells, DTD loses on all four backbones by 3–9 AUROC points, and wins scatter onto non-predicted datasets. AUROC agrees with AUOSCR, so text is the better *rejector*, not merely the better classifier. A residual signal remains, reported honestly: a per-area covariance variant beats a flat per-class Gaussian (GHOST) on the highest-resolution backbone (EVA-CLIP L/14-336) and gives small wins on a few strong backbones; local geometry carries *some* signal, just not robustly and never beating the global text score.

The two-sided answer, and its single cause. Both results share one mechanism. Patch-area matching sharpens closed-set discrimination because local texture and part patterns *transfer* across images of a class; that same transferability is fatal for rejection, since an unknown shares low-level texture with known classes, so its patches fall into known areas and coverage cannot flag it. Local features thus help closed-set *classification* but not open-set *rejection*, for the same reason.

9 Conclusion

We asked *when* local features help fine-grained recognition in VLMs. GL-VLM, a training-free few-shot patch-area descriptor, gives a cue-type-predictable gain (texture/part categories, neutral on global-shape) across 15 backbones and matches a *trained* adapter without training. It is a label-conditioned few-shot classifier (β=0 recovers the underlying global/Tip-Adapter baseline, not zero-shot CLIP), adjacent to cache/prototype adapters (Zhang et al., 2022) with novelty in the patch-region formulation, and does not address retrieval, generative, or open-set tasks. The contribution is the predictable dissociation and training-free parity, not state-of-the-art accuracy; because the backbone is never updated, the pretrained model and its zero-shot representation remain unchanged.

10 Use of AI Assistance

Per the ICLR 2027 AI policy, we disclose that large language model assistants were used (i) for retrieval of related work, and (ii) to aid writing. The architecture diagram (Figure 1) was drafted with AI assistance; all other figures and all reported numbers are the authors' own.

REFERENCES

Ioana Bica, Anastasija Ilić, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A. Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, and Jovana Mitrović. Improving fine-grained understanding in image-text pre-training. In *International Conference on Machine Learning (ICML)*, volume 235, pp. 3974–3995, 2024.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, 2014.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. arXiv:2407.01449.

Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos, Muhammad Jehanzeb Mirza, Margret Keuper, and Janis Keuper. Can we talk models into seeing the world differently? In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. arXiv:2403.09193.

Dipam Goswami, Simone Magistri, Gido M. van de Ven, Bartłomiej Twardowski, Andrew D. Bagdanov, Tinne Tuytelaars, and Joost van de Weijer. Cross-modal prototype alignment and mixing for training-free few-shot classification. *arXiv preprint arXiv:2603.24528*, 2026.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 39–48, 2020.

Savya Khosla, Sethuraman T V, Aryan Chadha, Alex Schwing, and Derek Hoiem. T-REN: Learning text-aligned region tokens improves dense vision-language alignment and scalability. *arXiv preprint arXiv:2604.18573*, 2026.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision (ICCV) Workshops*, pp. 554–561, 2013.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ProxyCLIP: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision (ECCV)*, 2024.

Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations (ICLR)*, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, 2022a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10965–10975, 2022b.

Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. SegCLIP: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning (ICML)*, volume 202, pp. 23033–23044, 2023.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Yifei Ming and Yixuan Li. How does fine-tuning impact out-of-distribution detection for vision-language models? *International Journal of Computer Vision (IJCV)*, 132(2):596–609, 2024.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, pp. 722–729, 2008.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Ryan Rabinowitz, Steve Cruz, Manuel Günther, and Terrance E. Boult. GHOST: Gaussian hypothesis open-set technique. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6666–6674, 2025. doi: 10.1609/aaai.v39i6.32715.

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 *International Conference on Machine Learning (ICML)*, 2021.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. SuS-X: Training-free name-only transfer of vision-language models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2725–2736, 2023.

Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision (ECCV)*, 2024a.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free CLIP-based adaptation. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024b.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic segmentation emerges from text supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18134–18144, 2022.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations (ICLR)*, 2022.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research (TMLR)*, 2022.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaptation of clip for few-shot classification. In *European Conference on Computer Vision (ECCV)*, 2022.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based language-image pretraining. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16793–16803, 2022.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision (ECCV)*, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 130(9), 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 (ICCV)*, 2023.