SELECTIVE AMBIGUITY REPAIR: PRESERVING CLIP'S PRIOR IN FEW-SHOT ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot adaptation should not alter what a pretrained model already gets right. However, most CLIP adaptation methods modify prompts, representations, or predictions globally, even when errors arise from confusion among only a few plausible classes. We formulate few-shot adaptation as *selective ambiguity repair*: preserve reliable pretrained decisions and intervene only where ambiguity is concentrated. We introduce a frozen-backbone framework that extracts complementary evidence across visual depth and applies bounded, ambiguity-aware residual corrections to a query-specific set of ambiguous classes. A complementary full-class branch provides a recovery path when the correct class falls outside this set, while adaptive fusion balances correction against preservation. Across 11 few-shot benchmarks, our method consistently improves strong protocol-matched CLIP adaptation baselines across 1/2/4/8/16 shots, with the gains extending consistently from RN50 to ViT-B/16. Ambiguity diagnostics further show that selective repair reduces candidate entropy, increases ground-truth margins, resolves more ambiguous predictions than dense or random correction, and better preserves reliable CLIP decisions. These results support a broader principle for few-shot vision–language adaptation: *concentrate adaptation capacity where the pretrained model is uncertain while preserving its reliable decision structure elsewhere.*

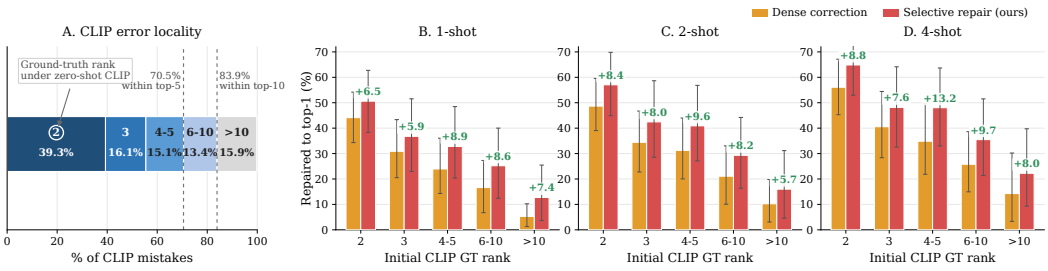

Figure 1: **Few-shot CLIP errors are predominantly local, motivating selective repair. A:** For zero-shot CLIP errors, the ground-truth class remains within the top-5 in 70.5% of cases and within the top-10 in 83.9%, indicating that failures are often concentrated within a small ambiguity set. **B–D:** We stratify errors by the initial ground-truth rank and report recovery to top-1 after $1/2/4$-shot adaptation. Candidate-selective repair consistently recovers more errors than dense correction across all rank groups, supporting the principle of correcting local ambiguity while preserving the remaining decision structure.

1 INTRODUCTION

Pretrained vision–language models such as CLIP Radford et al. (2021) transfer remarkably well to downstream recognition, yet adapting them from only a few labeled examples remains surprisingly delicate. With scarce supervision, the model must acquire task-specific discrimination without overwriting useful semantic structure learned at scale. This tension has motivated prompt learning Zhou et al. (2022b;a), adapters and cache-based methods Gao et al. (2024); Zhang et al. (2022), and more recent parameter-efficient or regularized adaptation approaches Bendou et al. (2025); Farina et al. (2025); Fahes et al. (2026). Despite their different implementations, most approaches modify prompts,

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.

representations, or predictions broadly. We ask a different question: *does few-shot adaptation need to intervene everywhere?*

Our key observation is that many CLIP errors are *local rather than global*. As illustrated in Fig. 1, among samples misclassified by zero-shot CLIP, the ground-truth class remains within the top-5 predictions in 70.5% of the cases and within the top-10 in 83.9%. CLIP therefore often identifies the correct semantic neighborhood, but fails to resolve a small set of plausible alternatives. This changes the adaptation problem. If most of the class space is already well organized, globally reshaping it from a handful of examples may be unnecessary and potentially harmful. Instead, adaptation should preserve reliable pretrained decisions and concentrate its capacity on the ambiguity that remains.

We formulate this principle as *Selective Ambiguity Repair* (SAR). For each query, SAR identifies a small set of competing classes and selectively modifies their relative evidence while preserving the surrounding CLIP decision structure. The challenge is twofold: determining *where* the ambiguity lies and determining *how* the competing classes should be separated. For the latter, we exploit information that is normally discarded when only CLIP's final visual embedding is used. Evidence that persists across visual depth provides stable semantic support, whereas changes between successive depths expose discriminative cues that emerge as the representation becomes more task-specific. We refer to these as *depth-consistent* and *depth-transition* evidence, respectively.

These complementary signals are fused to construct a CLIP-anchored prediction proposal, from which SAR predicts a bounded residual correction. Crucially, the residual intervention is applied only to a query-specific ambiguity set rather than to the full class distribution. A complementary full-class branch retains a recovery path when the correct class is not contained in that set, and shot-adaptive fusion controls the balance between preservation and correction as supervision increases. Thus, our claim is not that every computation is sparse: *the adaptation intervention itself is selective*. This distinction matters because it separates the proposed principle from dense logit correction and avoids perturbing classes where CLIP already provides reliable evidence.

Extensive experiments support both the effectiveness and the mechanism of this design. Across 11 standard recognition benchmarks and $1/2/4/8/16$-shot settings, SAR consistently improves strong protocol-matched RN50 adaptation baselines. The same mechanism also scales strongly to ViT-B/16, indicating that the benefit is not tied to a particular visual backbone. More importantly, direct ambiguity diagnostics show that SAR reduces candidate entropy, increases ground-truth margins, resolves more initially ambiguous predictions than dense or randomly targeted correction, and better preserves initially correct CLIP decisions. Multi-seed evaluation further shows that these improvements remain stable under different support-set samplings. Together, these results suggest that the main benefit does not arise from adding correction capacity alone but from *placing that capacity where the pretrained model is uncertain*.

Contributions. We make three contributions. First, we identify and quantify localized ambiguity as a dominant failure mode of few-shot CLIP, motivating adaptation as selective intervention rather than global modification. Second, we introduce SAR, which combines depth-consistent and depth-transition visual evidence with bounded, query-specific candidate repair while retaining an explicit full-class recovery path. Third, we provide broad empirical and mechanism-level validation across backbones, support-set samplings, distribution shifts, and ambiguity regimes, showing that selective repair improves the repair–preserve trade-off under scarce supervision.

2 Selective Ambiguity Repair (SAR)

Figure 2 presents the architecture of *Selective Ambiguity Repair* (SAR). Starting from a frozen CLIP backbone, SAR comprises three stages: *(i)* **depth-relational evidence fusion**, which extracts depth-consistent and depth-transition cues from intermediate visual features; *(ii)* **query-specific candidate repair**, which forms a CLIP-anchored full-class proposal, identifies a small ambiguity set, and applies a bounded residual only to the selected candidates; and *(iii)* **recovery and prediction fusion**, which combines the selectively repaired CLIP branch with a complementary full-class visual branch. The CLIP image and text encoders remain frozen throughout.

This decomposition separates three distinct roles that are easy to conflate: *evidence extraction* is class-complete, *candidate selection* uses the full score distribution to locate ambiguity, and only the subsequent *repair operator* is selective. The full-class branch is retained as a recovery path when the correct class falls outside the selected candidate set. Next, we formalize each component.

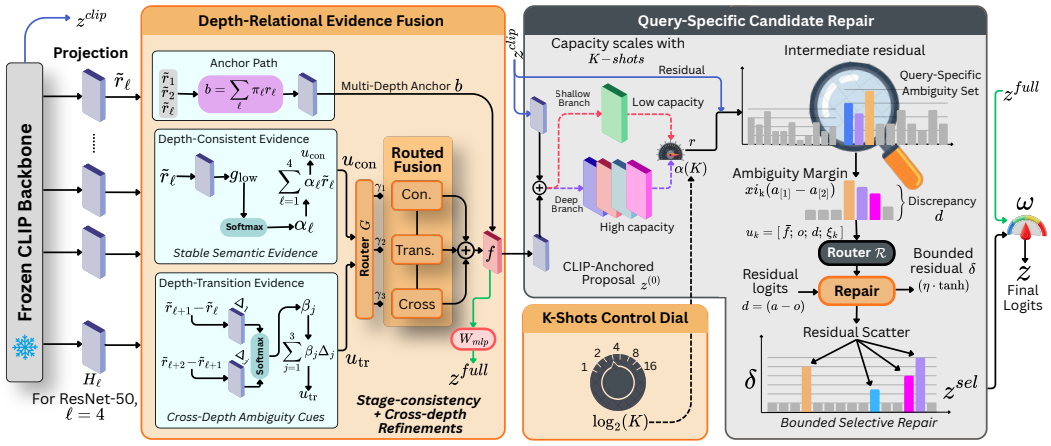

Figure 2: **Selective Ambiguity Repair (SAR).** Given a frozen CLIP backbone, SAR follows three stages. **(1) Depth-relational evidence fusion** combines *depth-consistent* evidence with *depth-transition* cues to obtain a fused representation f and a complementary full-class prediction z^{full}. **(2) Query-specific candidate repair** combines f with the frozen CLIP prior z^{clip} to form a CLIP-anchored proposal $z^{(0)}$, identifies its top-k ambiguity set, and applies a bounded residual only to those candidates, yielding z^{sel}. **(3) Recovery and prediction fusion** adaptively combines the selective branch with the full-class rec overy branch to produce the final prediction z. The full class space is used to *locate* ambiguity; only the selected candidates are *repaired*.

2.1 Problem Setup and CLIP Prior

We consider C-way K-shot classification with support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{CK}$. The pretrained CLIP image encoder f_v and text encoder f_t remain frozen. Let $T = [t_1, \ldots, t_C] \in \mathbb{R}^{d\times C}$ denote the normalized text prototypes constructed using the standard dataset-specific prompt templates, and let $v = \text{Norm}(f_v(x)) \in \mathbb{R}^d$ be the visual embedding of a query image x. The zero-shot CLIP logits are $z^{\text{clip}} = \tau T^\top v \in \mathbb{R}^C$, where τ is CLIP's pretrained logit scale. SAR uses z^{clip} as a semantic reference for both candidate selection and prior-preserving adaptation.

2.2 Depth-Relational Evidence Fusion

The final CLIP embedding may suppress intermediate visual cues that distinguish similar classes. We therefore extract representations $\{H_\ell\}_{\ell=1}^{L}$ from L depths of the frozen visual encoder and map them to a shared m-dimensional space, $r_\ell = \text{SSF}_\ell(\mathcal{P}_\ell(H_\ell)) \in \mathbb{R}^m, \quad \ell = 1, \ldots, L$, where $\mathcal{P}_\ell$ is a stage-specific lightweight projection and $\text{SSF}(x; \gamma, \beta) = \gamma \odot x + \beta$ is a learnable scale–shift transformation.

We seek two complementary evidence types. Information consistent across depth likely encodes stable CLIP-supported semantics. Changes between adjacent depths capture discriminative cues as representations are refined. We call these *depth-consistent* and *depth-transition* evidence.

We retain a conservative multi-depth anchor $b = \sum_{\ell=1}^{L} \pi_\ell r_\ell,\ \pi_\ell \geq 0,\ \sum_{\ell=1}^{L} \pi_\ell = 1$, where fixed weights $\{\pi_\ell\}$ mildly favor later stages while retaining earlier evidence. We then normalize $\tilde{r}_\ell = \text{LN}(r_\ell)$ and compute

$$\alpha_\ell = \text{softmax}_\ell\left(g_{\text{con}}(\tilde{r}_\ell)\right),\ u_{\text{con}} = \sum_{\ell=1}^{L} \alpha_\ell \tilde{r}_\ell,\ \beta_\ell = \text{softmax}_\ell\left(g_{\text{tr}}(\Delta_\ell)\right),\ u_{\text{tr}} = \sum_{\ell=1}^{L-1} \beta_\ell \Delta_\ell. \quad (1)$$

Here, $\Delta_\ell = \tilde{r}_{\ell+1} - \tilde{r}_\ell$. u_{con} summarizes evidence shared across depth, and u_{tr} captures how it changes between neighboring representations. This is not a spatial or Fourier-frequency decomposition, but one along representational depth.

Instead of treating both signals as equally useful for every query, we use a lightweight router with three role-defined experts: one for consistent evidence, one for transition evidence, and one for their

interaction. With $u = [u_{\text{con}}; u_{\text{tr}}]$ and $\gamma = \text{softmax}(G(u))$,

$$m = \gamma_{\text{con}} E_{\text{con}}(u_{\text{con}}) + \gamma_{\text{tr}} E_{\text{tr}}(u_{\text{tr}}) + \gamma_{\text{cross}} E_{\text{cross}}(u),\ f = b + \lambda m,\ z^{\text{full}} = W_{\text{full}} f. \tag{2}$$

The residual form $f = b + \lambda m$ is deliberate: the routed component can emphasize query-specific depth relations, but it refines rather than replaces the stable multi-depth anchor. The resulting $z^{\text{full}} \in \mathbb{R}^C$ provides a complementary full-class prediction that will later serve as a recovery path when candidate selection is imperfect.

2.3 Query-Specific Candidate Repair

The fused evidence f tells us *how* visually similar candidates should be separated, while z^{clip} tells us *where* CLIP currently places semantic probability mass. We combine the two to construct a CLIP-anchored proposal before applying any selective intervention.

Shot-adaptive proposal. The reliability of a learned correction should depend on the available supervision. We therefore define a generic shot-adaptive transformation

$$a(K) = a_{\min} + (a_{\max} - a_{\min})\sigma(\rho(\log_2 K - \mu)),\ \text{Mix}(x; K) = (1 - a(K))\psi_{\text{sh}}(x) + a(K)\psi_{\text{dp}}(x), \tag{3}$$

where ψ_{sh} and ψ_{dp} are low- and higher-capacity transformations with compatible outputs. Thus, adaptation favors the shallow path when supervision is scarce and smoothly increases its effective capacity as K grows. Let $\bar{z} = \text{SSF}_c(z^{\text{clip}})$ and $\bar{f} = \text{SSF}_f(f)$. We construct the full-class proposal

$$z^{(0)} = z^{\text{clip}} + \text{Mix}_c\left(\phi(\bar{z}) + \text{Mix}_f(\bar{f}; K); K\right). \tag{4}$$

$\phi(\cdot)$ is a lightweight projection that maps $\bar{z}$ to the same dimensionality as $\text{Mix}_f(\bar{f}; K)$ before fusion. The residual form keeps the proposal explicitly anchored to the frozen CLIP prediction. Importantly, $z^{(0)}$ is *not* the selective repair; it reveals the query's most likely competing classes.

Select, then repair. Let $\mathcal{I}_k = (i_1, \ldots, i_k)$ denote the indices of the top-k entries of $z^{(0)}$, where $k \ll C$ is a small fixed repair budget. We gather the proposal, original CLIP scores and define the candidate margin ξ_k as

$$a = z^{(0)}_{\mathcal{I}_k},\qquad o = z^{\text{clip}}_{\mathcal{I}_k},\qquad d = a - o,\quad \xi_k = \begin{cases} a_1 - a_2, & k > 1, \\ 0, & k = 1. \end{cases} \tag{5}$$

The repair head receives $u_k = [\,\bar{f};\, o;\, d;\, \xi_k\,]$, so that it jointly observes the visual evidence of the query, the original belief of CLIP, the proposed change from that belief, and the separation between the leading candidates. A lightweight routed head predicts a bounded residual only for these k classes,

$$\rho_k = \text{softmax}(R(u_k)),\qquad \delta = \eta \tanh\left(\sum_{e=1}^{E} \rho_{k,e} E_e(u_k)\right) \in \mathbb{R}^k, \tag{6}$$

where η bounds the maximum correction magnitude. We scatter δ back to the full class space,

$$[r^{\text{sel}}]_j = \begin{cases} \delta_t, & j = i_t,\ t \in \{1, \ldots, k\}, \\ 0, & j \notin \mathcal{I}_k, \end{cases} \qquad z^{\text{sel}} = z^{(0)} + r^{\text{sel}}. \tag{7}$$

$$\text{Hence the repair operator satisfies} \quad \text{supp}(r^{\text{sel}}) \subseteq \mathcal{I}_k,\quad |\mathcal{I}_k| = k \ll C. \tag{8}$$

Equation 8 defines selective ambiguity repair: all classes help locate the ambiguity, but the repair does not directly change non-candidate logits. Appendix Proposition 1 formalizes this locality and boundedness, and Corollary 1 shows the selective branch exactly preserves pairwise margins among non-candidate classes.

Table 1: **Average few-shot comparison.** Average accuracy (%) over 11 benchmarks using CLIP RN50. We show the top-10 strongest baselines; full per-dataset results are provided in Appendix B.1. The strongest baseline at each shot is underlined, and our result is highlighted.

Shot	Tip-F	TaskRes	GraphAdp.	GDA	ProKeR	TIMO-S	CCA-FT	LDC	SimNL	ProLIP	Ours
1-shot	64.6	64.3	64.8	62.2	64.0	66.5	66.0	65.7	<u>67.4</u>	64.2	**69.7**
2-shot	66.7	67.6	67.7	66.2	67.6	68.8	68.6	67.9	<u>70.0</u>	67.4	**72.5**
4-shot	69.7	70.3	70.3	69.8	70.7	71.6	72.1	71.2	<u>72.4</u>	70.5	**75.6**
8-shot	72.5	73.4	73.4	73.3	74.7	74.2	74.8	<u>75.8</u>	75.0	73.7	**78.7**
16-shot	75.7	75.8	76.2	76.0	76.8	76.9	77.6	<u>79.8</u>	77.8	76.5	**81.3**

2.4 Recovery, Prediction, and Learning

Candidate selection is intentionally conservative, but it can fail if the true class is outside $\mathcal{I}_k$. We therefore keep z^{full}, predicted directly from depth-relational visual evidence, as a recovery branch. Unlike selective repair, z^{sel} preserves and locally corrects the CLIP prior, while z^{full} supplies class-complete, task-specific evidence when the ambiguity set is incomplete. Their relative reliability also changes with the amount of supervision. We therefore compute a sample-wise shot-adaptive gate

$$\omega = \sigma((1 - a_g(K))s_{\text{sh}}(f) + a_g(K)s_{\text{dp}}(f)) \,,\; z = \omega z^{\text{full}} + (1 - \omega)z^{\text{sel}}, \tag{9}$$

where $a_g(K)$ follows the same form as Eq. 3 with independent parameters. The final logits z are used for inference. We supervise both branches individually, as well as their fused prediction:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{ce}}(z^{\text{full}}, y) + \lambda_2 \mathcal{L}_{\text{ce}}(z^{\text{sel}}, y) + \lambda_3 \mathcal{L}_{\text{ce}}(z, y) + \lambda_4 \left\| z^{\text{full}} - z^{\text{clip}} \right\|_1 + \lambda_5 \left\| z^{\text{sel}} - z^{\text{clip}} \right\|_1 . \tag{10}$$

Branch-level supervision prevents the final gate from compensating for a weak component, while the two prior-anchoring terms discourage unnecessary drift from CLIP's pretrained decision structure.

What is selective in SAR? SAR does not claim that all intermediate computations are sparse. The full-class proposal in Eq. 4 is required to identify the query's ambiguity set, and z^{full} deliberately remains class-complete to recover from candidate-selection errors. Selectivity refers specifically to the learned repair operator: r^{sel} has support only on $\mathcal{I}_k$. The model therefore separates *global evidence for deciding where to intervene* from *local intervention on the classes that require repair*.

2.5 Experimental Setup

Benchmarks and few-shot protocol. We evaluate SAR on the standard 11-dataset few-shot CLIP benchmark: ImageNet, Caltech101, OxfordPets, StanfordCars, Flowers102, Food101, FGVC-Aircraft, SUN397, DTD, EuroSAT, and UCF101. Following prior adapter- and cache-based methods, we report top-1 accuracy with $K \in \{1, 2, 4, 8, 16\}$ labeled examples per class. Our primary comparison uses CLIP RN50, which matches the baseline protocol in Table 1. Complete per-dataset results are provided in Appendix B, with implementation details in Appendix A.2.

Backbones and comparison scope. Our main few-shot comparison uses CLIP RN50. We additionally evaluate SAR with ViT-B/16 to assess transfer across backbone families and report a separate comparison with recent ViT-B/16 adaptation methods in Appendix C.2.

Generalization and robustness protocols. For base-to-novel (B2N) generalization, we follow the MaPLe protocol Khattak et al. (2023) using CLIP ViT-B/16: train on base classes, evaluate on base and unseen novel classes, and report Base, Novel, and their harmonic mean (HM). For distribution-shift robustness, we follow the LDC protocol Li et al. (2025a): SAR is adapted on ImageNet with RN50 and evaluated on ImageNet, ImageNet-V2, and ImageNet-Sketch. Each generalization comparison uses the backbone and evaluation protocol of its corresponding baseline family.

Robustness to few-shot sampling. Unless otherwise noted, we follow the support-set protocol used by the corresponding baselines. To quantify few-shot variability, we additionally rerun SAR and its closest dense logit-correction baseline, LDC, over five independently sampled support sets using identical splits, prompts, training schedules, and random-seed protocols. Complete per-dataset five-run statistics for both methods are provided in Appendix B.2.

The gains are stable across support sets. Table 2 shows that the improvement is not an artifact of a favorable few-shot sample. SAR outperforms LDC across all five shot regimes, with gains from

Table 2: **Robustness to support-set sampling.** Mean ± std over five independently sampled support sets across 11 datasets using RN50. Both methods use identical support sets and training protocols.

Method	1-shot	2-shot	4-shot	8-shot	16-shot
LDC	65.6±.56	68.3±.55	71.9±.54	76.0±.43	79.9±.16
SAR	**69.7±.22**	**72.5±.19**	**75.5±.22**	**78.5±.16**	**81.3±.15**
Δ	+4.1	+4.2	+3.6	+2.5	+1.4

Table 3: **Direct ambiguity resolution.** Metrics are averaged over $1/2/4/8/16$ shots on the fixed set of locally ambiguous frozen-CLIP errors. Lower H_k and higher M_y / Resolved are better.

Metric	Frozen CLIP	No repair	Random-k	Dense repair	Top-k SAR
$H_k \downarrow$	0.796	0.604	0.603	0.603	**0.593**
$M_y \uparrow$	−0.945	0.717	0.722	0.702	**0.733**
Resolved (%) ↑	0.00	59.27	59.45	59.32	**60.10**

$+1.4$ to $+4.2$ points and consistently lower run-to-run variation. More strongly, SAR's five-run mean exceeds LDC on all 55 dataset–shot combinations. The advantage therefore persists under changes in both support-set composition and optimization randomness.

2.6 Does Selective Repair Actually Resolve Ambiguity?

Accuracy alone cannot establish whether SAR works by resolving local ambiguity or simply by adding adaptation capacity. We therefore evaluate the prediction geometry directly. For samples misclassified by frozen CLIP whose ground-truth class lies within its original top-k candidates, we measure: (i) normalized candidate entropy H_k, where lower is better; (ii) the signed ground-truth margin $M_y = z_y - \max_{c \neq y} z_c$, where larger values indicate stronger separation; and (iii) the fraction of initially negative margins that become positive after adaptation (*Resolved*). The evaluation subset and candidate set are fixed using frozen CLIP before any adaptation.

SAR produces the lowest candidate entropy and the highest ambiguity-resolution rate at every shot level, while also achieving the strongest signed ground-truth margin on average (Table 3). Importantly, random-k repair uses the same k-class repair budget but does not match SAR, showing that the benefit is not sparsity alone: *where* the intervention is applied matters. Per-shot ambiguity diagnostics, together with ECE and NLL, are reported in Appendix D.2.

Selective repair improves the repair–preserve trade-off. Fig. 3 evaluates both sides of the proposed principle: correcting errors and preserving already-correct predictions. Relative to dense correction, SAR repairs $+6.8$, $+8.6$, and $+9.8$ percentage points more frozen-CLIP top-5 errors at $1/2/4$ shots while simultaneously preserving $+1.2$, $+1.6$, and $+0.7$ points more CLIP-correct predictions. Its score changes are also substantially more concentrated within CLIP's original top-ranked neighborhood. Thus, the improvement does not arise from simply changing logits more aggressively; *SAR concentrates its changes on fewer, more relevant classes*. Additional logit-locality analysis is provided in Appendix D.3. This empirical locality complements the formal preservation property in Appendix Proposition 1 and Corollary 1.

What happens when selection fails? A hard top-k intervention could appear brittle because the selective residual cannot directly increase the score of a ground-truth class outside the selected set. Table 4 tests this failure mode explicitly. While 63.6% of initial errors place the ground truth inside the selected candidate set, 36.4% do not. In the latter cases, the complementary full-class branch provides a genuine recovery path: when the ground truth lies within the top-10 but outside the selected set, recovery rises from 29.39% for the selectively repaired branch to 38.12% for the final model. Even for errors outside CLIP's top-10, final recovery increases from 10.56% to 15.06%. A more detailed analysis of this failure regime is provided in Appendix D.4.

The benefit is therefore not obtained by assuming perfect candidate selection. SAR combines a deliberately conservative local intervention with a class-complete recovery mechanism for cases in which the initial ambiguity set is incomplete.

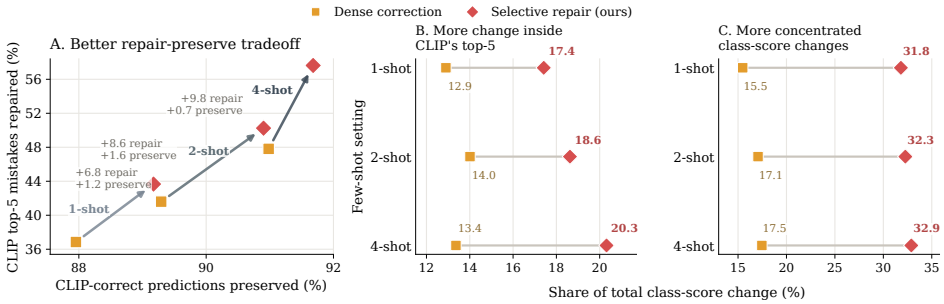

Figure 3: **Selective repair improves the repair–preserve trade-off by concentrating updates on ambiguous classes. A:** Repair–preserve analysis at 1/2/4 shots. Each arrow connects dense correction (square) to SAR (diamond); movement toward the upper right indicates simultaneously repairing more zero-shot CLIP errors while preserving more initially correct predictions. **B:** Fraction of total score change occurring within CLIP's original top-5 candidates. **C:** Fraction of total score change concentrated in the five most-updated classes. Across shot levels, SAR produces more localized interventions than dense correction while achieving a better balance between correcting errors and preserving reliable CLIP decisions.

Table 4: **Recovery when candidate selection is incomplete.** Recovery of initial frozen-CLIP errors across seven datasets and $1/4/16$ shots.

Initial GT location	Fraction	LDC	z^{sel}	Final SAR
Inside selected top-k	63.60	45.29	45.23	**50.04**
Top-10, outside top-k	23.01	31.18	29.39	**38.12**
Outside top-10	13.39	10.34	10.56	**15.06**

Qualitative behavior. Appendix Fig. 6 complements the aggregate analysis with representative predictions: in A and D, CLIP and LDC remain incorrect while SAR recovers the ground truth; in B, both adapted methods correct CLIP but SAR yields a larger ground-truth margin; and in C, CLIP is correct but uncertain, LDC flips the decision, whereas SAR preserves and strengthens it. Appendix Fig. 7 extends this analysis to broader operating regimes, including bounded correction and a failure case where the ground truth lies far outside CLIP's initial semantic neighborhood, while Fig. 8 examines fine-grained near-neighbor confusions. Together, these examples illustrate the behavior quantified above: SAR repairs local ambiguity while better preserving reliable pretrained decisions.

2.7 Generalization and Robustness

Base-to-novel generalization. Preserving CLIP's pretrained structure is especially important when adapting to one set of classes while maintaining transfer to unseen classes. We evaluate SAR under the standard MaPLe Khattak et al. (2023) base-to-novel protocol with CLIP ViT-B/16. Table 6 includes only methods in the standard few-shot B2N supervision setting; teacher-assisted methods such as PromptKD Li et al. (2024) and CAPT Shao et al. (2026), which use additional training resources, are discussed separately in Sec. 3. As shown in Table 6, SAR achieves the best harmonic mean (HM) in all six main-paper datasets, increasing the average HM from 72.9% to 75.0% (+2.2). The gain is not driven solely by base accuracy: SAR also improves novel-class accuracy in every dataset, with the highest HM increases on EuroSAT (+6.02), UCF101 (+2.70) and DTD (+2.57). The trend holds on the five additional benchmarks in Appendix C.1, where SAR again delivers the strongest HM on each dataset. In general, the results support the key idea of SAR: resolving task-specific ambiguities does not require a broad alteration of the semantic structure that benefits unseen classes.

Robustness under distribution shift. We test whether the gains persist under distribution shift. Following the ImageNet LDC protocol, we use RN50 on ImageNet and evaluate on ImageNet, ImageNet-V2, and ImageNet-Sketch. Table 7 shows that SAR achieves the best average accuracy (54.10%) vs. LDC (53.39%). SAR improves ImageNet (66.63% $\rightarrow$ 67.50%) and ImageNet-Sketch (35.52% $\rightarrow$ 36.90%), while remaining similar in ImageNet-V2 (57.90% vs. 58.03%). Thus, source-domain gains do not cause systematic OOD degradation, consistent with constrained adaptation rather than global perturbation of the pretrained decision space.

Table 5: **Core ablations of SAR. (a)** Top-k candidate repair outperforms dense and non-targeted alternatives, showing that the gain depends on *where* correction is applied. **(b)** Depth-consistent and depth-transition evidence are complementary, with routed fusion giving the strongest results. Average top-1 accuracy (%) over 11 datasets.

(a) Where should repair be applied?

Shot	No rep.	Dense	Rand.-k	Bottom-k	**Top-k**
1-shot	65.1	65.7	64.9	64.9	**69.7**
4-shot	71.4	71.2	71.1	71.4	**75.6**
16-shot	79.2	79.8	79.2	79.4	**81.3**

(b) What depth evidence is useful?

Shot	Uniform	Anchor	+Con.	+Trans.	+Both	**Full**
1-shot	63.6	64.6	64.8	65.0	66.4	**69.7**
4-shot	67.9	68.7	69.3	69.6	71.2	**75.6**
16-shot	75.4	76.8	77.3	77.5	78.7	**81.3**

Table 6: **Base-to-novel generalization under the MaPLe protocol Khattak et al. (2023).** Accuracy (%) on Base classes, Novel classes, and their harmonic mean (HM) across six datasets. The row below ours shows the difference from the strongest competing baseline for each metric. We achieve the best HM on all.

Method	(a) EuroSAT			(b) UCF101			(c) Caltech101			(d) DTD			(e) StanfordCars			(f) Aircraft		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	56.48	64.08	60.03	70.53	77.50	73.85	96.84	94.00	95.40	53.24	59.90	56.37	63.37	74.89	68.65	27.19	36.29	31.09
CoOp	91.54	54.44	68.27	85.14	64.47	73.37	98.00	89.81	93.73	79.44	41.18	54.24	**78.12**	60.40	68.13	40.44	22.30	28.75
CoCoOp	87.49	60.04	71.21	82.33	73.45	77.64	97.96	93.81	95.84	77.01	56.00	64.85	70.49	73.59	72.01	33.41	23.71	27.74
ProGrad	90.11	60.89	72.67	84.33	74.94	79.35	98.02	93.89	95.91	77.35	52.35	62.45	77.68	68.63	72.88	**40.54**	27.57	32.82
KgCoOp	85.64	64.34	73.48	82.89	76.67	79.65	97.72	94.39	96.03	77.55	54.99	64.35	71.76	75.04	73.36	36.21	33.55	34.83
MaPLe	94.07	73.23	82.35	83.00	78.66	80.77	97.74	94.36	96.02	80.36	59.18	68.16	72.94	74.00	73.47	37.44	35.61	36.50
Ours	**95.48**	**82.26**	**88.37**	**87.28**	**79.98**	**83.47**	**98.26**	**94.43**	**96.31**	**86.23**	**59.95**	**70.73**	72.99	**75.32**	**74.13**	36.85	**37.55**	**37.20**
	+1.41	+9.03	+6.02	+2.14	+1.32	+2.70	+0.24	+0.04	+0.28	+5.87	+0.05	+2.57	-5.13	+0.28	+0.66	-3.69	+1.26	+0.70

Transfer across visual backbones. Finally, we test whether selective repair depends on the RN50 hierarchy. Replacing RN50 with ViT-B/16 and extracting evidence from intermediate transformer blocks yields $76.2/78.7/80.7/82.4/84.4\%$ average accuracy at $1/2/4/8/16$ shots. Consistently strong results indicate selective repair transfers from a convolutional to a transformer-based CLIP visual encoder. See Appendix C.2 for comparison to recent ViT-B/16 adaptation methods.

2.8 Ablation Studies

Next, we isolate the impact of SAR's main design choices. Unless noted otherwise, results are averaged across 11 few-shot benchmarks at 1/4/16 shots using RN50.

Where should adaptation intervene? Table 5(a) isolates the effect of repair location. Applying a residual densely over all classes provides only modest improvement, while random-k and bottom-k correction are comparable to or worse than removing the repair module altogether. In contrast, targeting the query-specific top-k ambiguity set yields $69.7/75.6/81.3\%$ at $1/4/16$ shots, outperforming dense correction by $+4.0/+4.4/+1.5$ points. Thus, the benefit is not explained by residual capacity or sparsity alone: *which classes are allowed to change is critical*.

What visual evidence is needed for repair? Table 5(b) tests the depth-relational formulation. Uniform multi-depth aggregation is weakest, while retaining a stable multi-depth anchor already improves performance. Adding either depth-consistent or depth-transition evidence gives further gains, and combining them is consistently stronger than either alone. The full routed formulation performs best at every shot level, reaching $69.7/75.6/81.3\%$. These results support the intended decomposition: depth-consistent evidence preserves semantics shared across the visual hierarchy, while depth-transition evidence captures complementary discriminative changes; sample-dependent routing determines how strongly each should influence the query.

Computational cost. SAR's selectivity changes where predictions are modified, not whether class-complete evidence is extracted. We compare its cost to the closest dense logit-correction baseline, LDC, under the same ImageNet 16-shot RN50 setting on one NVIDIA A40. SAR uses 22.63M trainable parameters and 24.27 GFLOPs/image vs. 17.26M and 19.86 GFLOPs for LDC (+31.2% parameters, +22.2% compute). Peak allocated memory at batch size 64 is essentially unchanged for training (3.89 vs. 3.85 GB) and inference (2.22 vs. 2.20 GB). Thus, SAR adds moderate compute for multi-depth evidence extraction and selective repair with negligible GPU-memory overhead. Details and the measurement protocol are in Appendix F.

Table 7: **OOD evaluation under the ImageNet protocol, following LDC Li et al. (2025a). Avg** is the mean over the three datasets. Best results are in **bold**; our method is shaded.

Dataset	CLIP	CLAP	CoOp	ProGrad	GraphAd.	Tip-F	CLIP-Ad.	CCA-FT	TaskRes	ProKeR	LDC	Ours
ImageNet	60.33	65.02	62.59	62.17	64.94	65.51	61.33	66.04	65.73	64.66	66.63	**67.50**
ImageNet-V2	53.27	56.09	55.11	54.70	56.58	57.11	52.67	57.32	57.00	56.11	**58.03**	57.90
ImageNet-Sketch	35.44	34.55	32.74	34.40	35.89	36.00	32.04	35.87	34.43	36.08	35.52	**36.90**
Avg	49.35	51.89	46.81	50.42	52.47	52.87	48.68	53.08	52.39	52.28	53.39	**54.10**

3 Related Work

Few-shot adaptation of CLIP. Few-shot CLIP adaptation has been approached through prompt learning, feature-space tuning, and cache-based refinement. CoOp and CoCoOp Zhou et al. (2022b;a) learn textual contexts, while subsequent methods such as ProGrad, KgCoOp, and MaPLe Khattak et al. (2023) improve generalization or extend adaptation across modalities. Adapter- and cache-based approaches instead modify visual representations or predictions, including CLIP-Adapter Gao et al. (2024), Tip-Adapter Zhang et al. (2022), TaskRes Yu et al. (2023), and more recent structured or regularized variants such as ProKeR Bendou et al. (2025) and ProLIP Fahes et al. (2026). Recent methods also explicitly address preservation of pretrained knowledge: MMRL Guo & Gu (2025) separates task-specific and transferable representations, while PromptKD Li et al. (2024) transfers knowledge from a larger teacher using additional unlabeled target-domain data. These approaches differ substantially in supervision and adaptation scope, but share a central challenge: improving task-specific discrimination without eroding CLIP's transferable prior. SAR addresses this challenge at the decision level, asking not only *how* to adapt, but *which class decisions actually need to change*.

Structured visual evidence and decision-space correction. Several methods show that CLIP's final global embedding does not expose all information useful for few-shot recognition. GraphAdapter Li et al. (2023) and ProKeR Bendou et al. (2025) exploit finer-grained visual or relational structure, while LDC Li et al. (2025a) uses intermediate visual evidence to predict residual corrections directly in logit space. These results motivate both multi-depth evidence extraction and decision-level adaptation. However, multi-level evidence is typically aggregated as additional support, and logit-space correction is applied over the full class distribution. SAR instead distinguishes evidence that remains consistent across visual depth from evidence revealed by transitions between depths, and uses the resulting representation to determine a bounded, query-specific correction. Thus, the full class space is used to locate ambiguity, but only the selected candidates are repaired.

Confusion-aware adaptation. CAPT Shao et al. (2026) shares a similar motivation: it targets persistent confusion among visually/semantically similar classes by modeling it with a Confusion Bank, semantic difference/commonality prompts, and representative confusing samples. SAR tackles the same issue differently. CAPT learns reusable dataset-level class-pair and sample-level confusion structure within a PromptKD-based adaptation pipeline; SAR detects ambiguity per query from frozen CLIP predictions and applies a bounded residual only to selected candidate logits. In short, CAPT asks *how to learn and exploit recurring confusion structure*, while SAR asks *which decisions to change for the current query while preserving the rest of the pretrained prediction*.

4 Conclusion and Limitations

We introduce *Selective Ambiguity Repair* (SAR), motivated by the observation that many few-shot CLIP errors arise from uncertainty among a few plausible classes rather than failure of the entire pretrained decision space. SAR adapts by deciding *where to intervene*: depth-consistent and depth-transition evidence form a CLIP-anchored proposal, followed by a bounded residual on a query-specific ambiguity set and a complementary full-class recovery path. Across few-shot recognition, base-to-novel transfer, and distribution shift, SAR improves adaptation while preserving reliable pretrained decisions. Repair–preserve and ambiguity analyses further show that the gains come from *targeted* intervention rather than residual capacity or sparsity alone. More broadly, adaptation should refine a reliable semantic neighborhood rather than restructure the full prediction space.

Limitations. SAR depends on CLIP's pretrained ranking. If the true class is far from the initial semantic neighborhood, candidate-selective repair cannot elevate it, though the full-class branch partly recovers. The method also assumes closed-set classification with a fixed candidate budget and adds moderate compute from multi-depth evidence extraction. Future work includes dynamic or uncertainty-aware candidate sets, and open-vocabulary recognition.

REFERENCES

Yassir Bendou, Amine Ouasfi, Vincent Gripon, and Adnane Boukhayma. Proker: A kernel perspective on few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25092–25102, 2025.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Dexia Chen, Qianjie Zhu, Weibing Li, Yue Yu, Tong Zhang, and Ruixuan Wang. Preserve and sculpt: Manifold-aligned fine-tuning of vision-language models for few-shot learning. In *International Conference on Learning Representations*, pp. 109043–109065, 2026.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul De Charette. CLIP's visual embedding projector is a few-shot cornucopia. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3254–3264, 2026.

Matteo Farina, Massimiliano Mancini, Giovanni Iacca, and Elisa Ricci. Rethinking few-shot adaptation of vision-language models in two stages. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29989–29998, 2025.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

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

Yuncheng Guo and Xiaodong Gu. Mmrl: Multi-modal representation learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25015–25025, 2025.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. Lp++: A surprisingly strong linear probe for few-shot clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23773–23782, 2024.

Tianjiao Jiang, Zhen Zhang, Yuhang Liu, and Javen Qinfeng Shi. Causal disentanglement and cross-modal alignment for enhanced few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 890–900, 2025.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19113–19122, 2023.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Marc Lafon, Elias Ramzi, Clément Rambour, Nicolas Audebert, and Nicolas Thome. Gallop: Learning global and local prompts for vision-language models. In *European Conference on Computer Vision*, pp. 264–282. Springer, 2024.

Shuo Li, Fang Liu, Zehua Hao, Xinyi Wang, Lingling Li, Xu Liu, Puhua Chen, and Wenping Ma. Logits deconfusion with clip for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25411–25421, 2025a.

Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. *Advances in Neural Information Processing Systems*, 36:13448–13466, 2023.

Yayuan Li, Jintao Guo, Lei Qi, Wenbin Li, and Yinghuan Shi. Text and image are mutually beneficial: Enhancing training-free few-shot classification with clip. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5039–5047, 2025b.

Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26607–26616. IEEE, 2024.

Liangchen Liu, Nannan Wang, Xi Yang, Xinbo Gao, and Tongliang Liu. Surrogate prompt learning: Towards efficient and diverse prompt learning for vision-language models. In *ICML*, 2025.

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

Mohammed Rahman Sherif Khan Mohammad, Ardhendu Behera, Sandip Pradhan, Swagat Kumar, and Amr Ahmed. Training-only heterogeneous image-patch-text graph supervision for advancing few-shot learning adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19613–19622, June 2026.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

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

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Maoyuan Shao, Yutong Gao, Xinyang Huang, Lijuan Sun, Guoshun Nan, and Chuang Zhu. Capt: Confusion-aware prompt tuning for reducing vision-language misalignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3154–3164, 2026.

Julio Silva-Rodriguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23681–23690, 2024.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019.

Ran Wang, Hua Zuo, Zhen Fang, and Jie Lu. Integrated image-text augmentation for few-shot learning in vision-language models. *ACM Transactions on Intelligent Systems and Technology*, 16 (2):1–19, 2025.

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)*, 2024.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23826–23837. IEEE, 2024.

Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10899–10909, 2023.

Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1593–1603, 2024.

Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Enhancing vision-language few-shot adaptation with negative learning. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5905–5915. IEEE, 2025.

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

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022a.

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

AI USE STATEMENT

The authors used AI tools only for writing and editorial support during the preparation of the manuscript. Their use was limited to improving readability, presentation, and formatting. The research itself, including the development of the approach, the experimental work, the analysis of the results, and the interpretation of the findings, was carried out independently by the authors. The authors reviewed the final manuscript in full and remain responsible for its accuracy and content.

A EXPERIMENTAL PROTOCOLS AND IMPLEMENTATION

This section specifies the evaluation protocols and implementation choices underlying the experiments in the main paper. Complete per-dataset few-shot results are reported in Appendix B; generalization and backbone-transfer results in Appendix C; and additional mechanism-level analyses and ablations in Appendices D and E, respectively.

A.1 EVALUATION PROTOCOLS

Few-shot classification. We evaluate on the standard 11-dataset CLIP few-shot benchmark: ImageNet Deng et al. (2009), SUN397 Xiao et al. (2010), FGVC-Aircraft Maji et al. (2013), EuroSAT Helber et al. (2019), StanfordCars Krause et al. (2013), Food101 Bossard et al. (2014), OxfordPets Parkhi et al. (2012), Flowers102 Nilsback & Zisserman (2008), Caltech101 Fei-Fei et al. (2004), DTD Cimpoi et al. (2014), and UCF101 Soomro et al. (2012). For each dataset, we use $K \in \{1, 2, 4, 8, 16\}$ labeled examples per class and report top-1 classification accuracy. The support set is used only to optimize the lightweight SAR modules; the pretrained CLIP image and text encoders remain frozen. Our primary comparison uses CLIP RN50, which matches the protocol of the baselines reported in the main paper. To quantify sensitivity to support-set sampling, SAR and LDC are also evaluated over five independently sampled support sets using matched splits, prompts, training schedules and random-seed protocols; complete per-dataset statistics are reported in Appendix B.

Base-to-novel generalization. We follow the standard MaPLe protocol Khattak et al. (2023) with CLIP ViT-B/16. Each dataset is partitioned into base and novel classes; adaptation uses only labeled examples from the base classes, while evaluation is performed on both base and unseen novel classes. We report Base accuracy, Novel accuracy, and their harmonic mean (HM). This protocol tests whether task-specific adaptation can improve discrimination without sacrificing CLIP's transferable knowledge. Results on the remaining five benchmarks are provided in Appendix C.1, completing the 11-dataset evaluation together with Table 6.

Distribution-shift robustness. Following LDC Li et al. (2025a), we adapt CLIP RN50 on ImageNet Deng et al. (2009) and evaluate the resulting model without further adaptation on ImageNet, ImageNet-V2 Recht et al. (2019), and ImageNet-Sketch Wang et al. (2019). We report top-1 accuracy on each dataset and their average. This experiment evaluates whether improvements on the adaptation distribution persist under natural and stylistic distribution shifts.

A.2 IMPLEMENTATION DETAILS

Unless noted, few-shot experiments use CLIP RN50; ViT-B/16 is evaluated separately to test cross-backbone transfer. For both, pretrained CLIP image and text encoders are frozen. Dataset-specific prompt templates follow the protocols of the compared methods.

SAR trains only the lightweight components introduced in Sec. 2: stage-specific projections and scale–shift transformations, depth-relational evidence fusion, shot-adaptive proposal modules, the query-specific repair router and experts, the full-class recovery classifier, and the final prediction gate. For RN50, evidence comes from four residual stages; for ViT-B/16, from four intermediate transformer blocks. The selective repair operator updates only the k candidates in the query-specific ambiguity set, while full-class recovery remains class-complete.

Unless specified otherwise, we use a shared hyperparameter configuration across datasets for each evaluation protocol. Hyperparameters are tuned without access to test labels, novel-class labels, or

OOD target-set labels. Optimization settings and architecture-specific hyperparameters are summarized in Appendix A.3.

A.3 Training Configuration

Table 8 summarizes the optimization and architecture-specific settings used in the RN50 experiments. These values are held fixed across datasets within the corresponding evaluation protocol unless stated otherwise.

Table 8: **Training and architecture configuration for RN50.** Optimization settings and principal SAR hyperparameters used in the standard few-shot experiments.

Backbone	Opt.	LR	Batch	Epochs	WD	Fusion Dim.	k	Experts	Expert Dim.	Res. Init.
RN50	AdamW	1×10^{-3}	64	50	1×10^{-4}	256	5	3	256	0.05

Reproducibility. All experiments use public benchmarks and CLIP backbones. For each evaluation protocol, hyperparameters are fixed across datasets unless noted, and no test, OOD target, or novel-class labels are used for model selection. CLIP image/text encoders stay frozen; only lightweight SAR modules are optimized.

Support-set sampling. The main few-shot tables follow the baselines' support-set protocol. To test sensitivity to support composition, we also evaluate SAR and LDC on five independently sampled, matched support sets; full results are in Appendix B.2.

B Full Few-Shot Results and Statistical Robustness

This section extends the main paper's aggregate few-shot results in two ways. First, we report the full per-dataset RN50 comparison behind Table 1 to check whether average gains hold across datasets and shot regimes. Second, we present the complete five-run LDC comparison under matched support-set sampling, complementing the main paper's aggregate robustness results.

B.1 Per-Dataset RN50 Comparison

Tables 9 and 10 report the per-dataset results behind the aggregate RN50 comparison in Table 1. SAR achieves the highest cross-dataset average at all shot levels, outperforming the strongest baseline by $+2.3$, $+2.5$, $+3.2$, $+2.9$, and $+1.5$ points at $1/2/4/8/16$ shots. At 1-shot, SAR leads on 9 of 11 datasets, and its advantage becomes broadly positive as supervision increases.

The gains are not confined to one benchmark family. Margins are largest on EuroSAT, FGVC-Aircraft, DTD, Flowers102, and UCF101, and smaller on ImageNet, Food101, and OxfordPets where baselines are already strong. This matches SAR's motivation: it helps most when CLIP offers useful semantic structure but added task-specific discrimination is needed. We do not treat this dataset-level pattern as standalone evidence of localized ambiguity; the direct ambiguity and repair–preserve analyses provide mechanistic validation.

At 8 and 16 shots, dense logit correction becomes more competitive, yet SAR retains the highest average accuracy. Selective intervention is therefore not only a low-shot regularizer: even with stronger supervision, preserving reliable parts of the pretrained decision structure while concentrating correction on the competing classes remains beneficial.

B.2 Robustness to Support-Set Sampling

The main paper reports the cross-dataset five-run comparison in Table 2. Here we provide the complete per-dataset statistics. SAR and LDC are evaluated on the same five independently sampled support sets using matched splits, prompts, optimization schedules, and random-seed protocols.

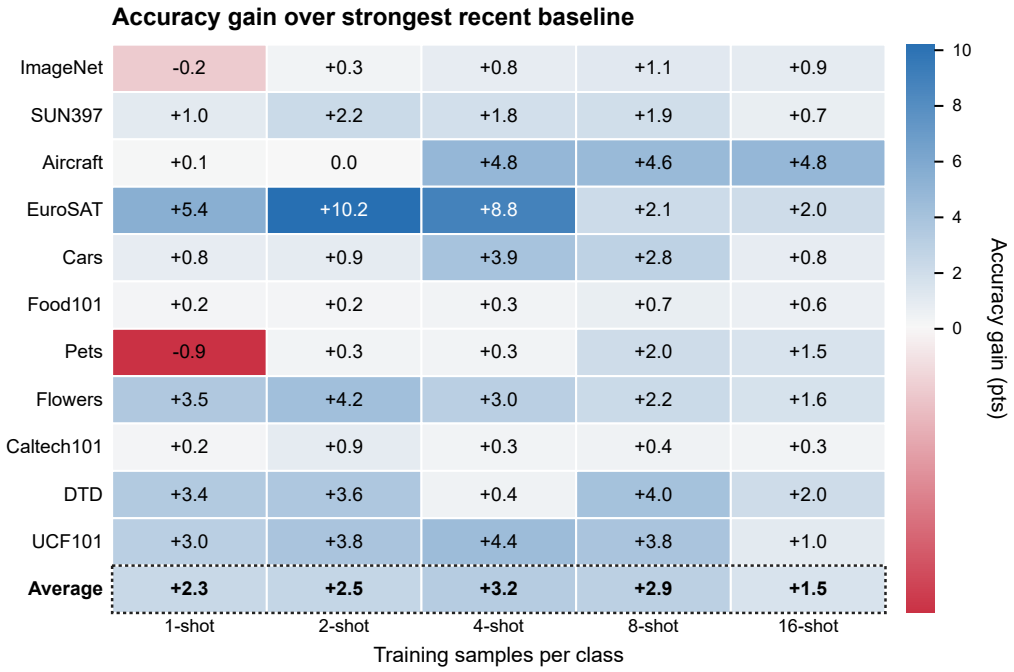

Figure 4: **Per-dataset gain over the strongest included protocol-matched RN50 baseline.** Each cell reports SAR's accuracy difference, in percentage points, from the strongest competing method in Tables 9 and 10 for the corresponding dataset and shot setting. The final row gives the cross-dataset improvement: $+2.3/+2.5/+3.2/+2.9/+1.5$ points from 1 to 16 shots. The broad positive pattern shows that the aggregate gain is not driven by a small subset of benchmarks.

Table 9: **Full protocol-matched RN50 comparison at 1/2/4 shots.** Top-1 accuracy (%) on the 11 few-shot benchmarks. The main paper reports cross-dataset averages; here we provide the complete per-dataset breakdown. The best result within each dataset and shot setting is shown in **bold**.

	Method	Venue	INet	SUN	Air	Euro	Cars	Food	Pets	Flow	Cal	DTD	UCF	Avg
0	CLIP	ICML'21	60.3	58.5	17.1	37.5	55.7	77.3	89.1	66.0	85.9	42.2	61.5	59.2
1	TIP-Adapter-F Zhang et al. (2022)	ECCV'22	61.1	62.5	20.2	59.5	58.9	77.5	87.0	80.0	89.3	49.6	64.9	64.6
	TaskRes Yu et al. (2023)	CVPR'23	61.9	62.3	21.4	61.7	59.1	74.0	83.6	79.2	88.8	50.2	64.8	64.3
	GraphAdapter Li et al. (2023)	NeurIPS'23	61.5	61.9	20.9	63.3	59.7	75.4	84.4	80.0	88.9	51.8	64.9	64.8
	CLIP-Adapter Gao et al. (2024)	IJCV'24	61.2	61.3	17.4	61.4	55.1	76.8	85.9	73.4	88.6	45.8	62.2	62.6
	CLAP Silva-Rodriguez et al. (2024)	CVPR'24	58.5	61.1	20.6	59.2	56.3	73.0	83.6	79.9	88.4	47.5	62.5	62.8
	GDA Wang et al. (2024)	ICLR'24	60.6	60.0	17.3	58.3	56.8	77.4	85.5	72.1	87.3	46.3	62.6	62.2
	LP++ Huang et al. (2024)	CVPR'24	61.1	62.4	19.6	57.2	57.2	76.6	84.2	78.2	88.5	46.9	65.4	63.4
	CCA-FT Jiang et al. (2025)	ICCV'25	61.4	63.5	21.6	67.1	59.7	77.6	86.8	81.0	89.8	51.0	66.0	66.0
	ProKeR Bendou et al. (2025)	CVPR'25	60.5	61.5	20.5	60.0	59.0	77.6	86.1	79.0	89.0	48.5	64.5	64.0
	TIMO-S Li et al. (2025b)	AAAI'25	62.2	65.1	23.5	62.1	59.8	77.3	**88.4**	83.0	89.6	54.3	67.1	66.5
	LDC Li et al. (2025a)	CVPR'25	60.5	62.5	18.7	73.9	57.0	76.4	87.2	80.0	89.7	50.1	66.8	65.7
	SimNL Zhang et al. (2025)	WACV'25	**62.8**	65.1	23.6	67.7	61.3	77.9	86.9	84.1	90.8	53.1	68.2	67.4
	ProLIP Fahes et al. (2026)	WACV'26	61.3	63.4	19.7	57.9	58.7	77.0	85.4	75.3	89.2	50.2	67.8	64.2
	Ours	—	62.6	**66.1**	**23.7**	**79.3**	**62.1**	**78.1**	87.5	**87.6**	**91.0**	**57.7**	**71.2**	**69.7**
2	TIP-Adapter-F Zhang et al. (2022)	ECCV'22	61.7	63.6	23.2	66.2	61.5	77.8	87.0	82.3	89.7	53.7	66.4	66.7
	TaskRes Yu et al. (2023)	CVPR'23	62.6	65.0	24.1	65.8	63.7	75.2	84.6	86.6	90.3	55.1	70.0	67.6
	GraphAdapter Li et al. (2023)	NeurIPS'23	62.3	64.6	23.8	67.3	63.2	76.3	86.3	85.6	90.2	55.7	69.5	67.7
	CLIP-Adapter Gao et al. (2024)	IJCV'24	61.5	63.3	20.1	63.9	58.7	77.2	86.7	81.6	89.4	51.5	67.1	65.5
	CLAP Silva-Rodriguez et al. (2024)	CVPR'24	58.5	63.3	23.2	65.6	61.4	74.9	84.9	84.2	89.8	53.1	67.8	66.1
	GDA Wang et al. (2024)	ICLR'24	61.2	63.6	22.4	67.9	59.3	77.9	86.7	82.9	88.6	51.3	66.3	66.2
	LP++ Huang et al. (2024)	CVPR'24	61.5	64.6	21.5	61.6	59.9	77.2	85.7	84.6	89.5	52.4	69.2	66.2
	CCA-FT Jiang et al. (2025)	ICCV'25	62.2	66.3	25.2	69.9	63.1	77.8	87.6	87.4	90.9	55.0	68.9	68.6
	ProKeR Bendou et al. (2025)	CVPR'25	61.4	64.3	23.7	70.0	61.2	77.9	87.3	86.8	89.7	54.2	68.8	67.6
	TIMO-S Li et al. (2025b)	AAAI'25	62.6	66.7	24.5	69.1	61.7	78.0	88.0	86.8	91.4	57.8	70.4	68.8
	LDC Li et al. (2025a)	CVPR'25	61.3	64.5	21.5	74.5	59.8	77.2	87.1	86.5	90.2	54.7	69.8	67.9
	SimNL Zhang et al. (2025)	WACV'25	63.4	66.7	26.2	74.4	64.7	78.3	87.6	88.4	91.1	59.1	70.2	70.0
	ProLIP Fahes et al. (2026)	WACV'26	61.7	65.1	22.6	70.0	61.7	77.6	86.1	81.9	89.8	54.7	70.0	67.4
	Ours	—	**63.7**	**68.9**	**26.3**	**84.7**	**65.6**	**78.5**	**88.3**	**92.6**	**92.4**	**62.7**	**74.4**	**72.6**
4	TIP-Adapter-F Zhang et al. (2022)	ECCV'22	62.5	66.2	25.8	74.1	64.6	78.2	87.5	88.8	90.6	57.4	70.6	69.7
	TaskRes Yu et al. (2023)	CVPR'23	63.6	67.3	25.7	73.8	67.4	76.1	86.3	90.2	91.0	60.7	70.9	70.3
	GraphAdapter Li et al. (2023)	NeurIPS'23	63.1	66.7	27.0	75.2	66.5	76.8	86.6	89.9	91.0	59.6	71.5	70.3
	CLIP-Adapter Gao et al. (2024)	IJCV'24	61.8	65.9	22.5	73.3	62.4	77.9	87.4	87.1	89.9	56.8	69.0	68.6
	CLAP Silva-Rodriguez et al. (2024)	CVPR'24	60.7	65.9	25.6	73.1	65.5	75.9	86.5	87.6	90.6	58.8	69.8	69.1
	GDA Wang et al. (2024)	ICLR'24	61.7	65.3	28.2	75.4	62.8	77.8	87.2	89.7	90.6	57.5	71.3	69.8
	LP++ Huang et al. (2024)	CVPR'24	62.5	67.2	24.2	68.6	63.4	77.7	86.9	89.5	90.8	57.7	71.6	69.1
	CCA-FT Jiang et al. (2025)	ICCV'25	63.3	68.9	28.6	80.4	67.6	78.2	88.2	91.2	92.0	62.4	71.7	72.1
	ProKeR Bendou et al. (2025)	CVPR'25	62.1	66.5	28.7	78.6	65.0	78.2	87.8	90.1	90.9	59.7	72.8	70.7
	TIMO-S Li et al. (2025b)	AAAI'25	63.3	68.5	29.9	75.2	65.9	78.4	88.7	91.6	91.7	61.2	72.7	71.6
	LDC Li et al. (2025a)	CVPR'25	62.5	67.8	26.4	80.5	64.8	77.8	88.1	91.6	90.8	59.3	73.2	71.2
	SimNL Zhang et al. (2025)	WACV'25	64.1	68.5	28.9	76.5	67.9	78.7	88.1	92.0	92.2	66.0	73.4	72.4
	ProLIP Fahes et al. (2026)	WACV'26	62.3	67.3	27.0	76.4	65.6	77.7	87.0	88.3	91.4	59.3	73.5	70.5
	Ours	—	**64.9**	**70.7**	**34.7**	**89.3**	**71.8**	**79.0**	**89.2**	**95.0**	**92.5**	**66.4**	**77.9**	**75.6**

Table 11 shows that SAR's five-run mean exceeds LDC on all 55 dataset–shot combinations. The advantage is largest under scarce supervision but remains positive through 16 shots. At the cross-dataset level, SAR improves LDC by $+4.1/+4.2/+3.6/+2.5/+1.4$ points at $1/2/4/8/16$ shots. Its run-to-run variation is also lower for the aggregate accuracy at 1–8 shots and comparable at 16 shots. These results rule out a favorable support-set draw as the explanation for the improvements observed in the standard benchmark comparison.

C Generalization and Backbone Transfer

The main paper evaluates whether SAR preserves transferable CLIP structure through base-to-novel generalization and ImageNet distribution shift (Sec. 2.7). Here we complete the base-to-novel evaluation on the remaining five benchmarks and provide a separate ViT-B/16 few-shot comparison against recent adaptation methods. The full OOD comparison is already reported in Table 7 of the main paper and is not duplicated here.

Table 10: **Full protocol-matched RN50 comparison at 8/16 shots.** Top-1 accuracy (%) on the 11 few-shot benchmarks. The best result within each dataset and shot setting is shown in **bold**. SAR retains the strongest cross-dataset average as supervision increases.

	Method	Venue	INet	SUN	Air	Euro	Cars	Food	Pets	Flow	Cal	DTD	UCF	Avg
8	TIP-Adapter-F	ECCV'22	64.0	68.9	30.2	77.9	69.2	78.6	88.1	91.5	91.4	62.7	74.2	72.5
	TaskRes Yu et al. (2023)	CVPR'23	64.7	68.7	31.5	79.3	71.8	76.4	87.2	94.7	92.4	64.8	75.3	73.4
	GraphAdapter Li et al. (2023)	NeurIPS'23	64.2	68.9	31.4	80.2	70.5	77.7	87.6	94.1	92.4	64.5	75.7	73.4
	CLIP-Adapter Gao et al. (2024)	IJCV'24	62.7	67.5	26.2	77.9	67.9	78.0	87.7	91.7	91.4	61.0	73.3	71.8
	CLAP Silva-Rodriguez et al. (2024)	CVPR'24	63.0	68.6	29.0	76.7	70.4	77.4	87.8	92.1	91.5	63.2	73.3	72.1
	GDA Wang et al. (2024)	ICLR'24	62.5	68.6	34.1	81.7	68.9	78.4	89.1	93.2	91.3	62.8	75.7	73.3
	LP++ Huang et al. (2024)	CVPR'24	63.7	69.3	27.7	75.8	67.8	78.5	87.7	92.6	91.8	62.4	74.8	72.0
	CCA-FT Jiang et al. (2025)	ICCV'25	64.7	70.8	35.3	82.4	73.3	79.2	89.0	93.8	92.5	64.9	77.3	74.8
	ProKeR Bendou et al. (2025)	CVPR'25	63.6	69.0	34.8	80.8	70.5	78.7	88.4	92.4	92.0	65.0	75.3	74.7
	TIMO-S Li et al. (2025b)	AAAI'25	64.0	70.2	34.7	81.8	69.8	78.5	89.2	94.4	92.6	65.1	75.2	74.2
	LDC Li et al. (2025a)	CVPR'25	64.4	70.6	39.4	87.9	74.4	79.0	89.0	93.6	92.6	65.8	76.9	75.8
	SimNL Zhang et al. (2025)	WACV'25	65.3	70.9	33.9	81.6	72.0	79.2	89.2	95.2	93.4	67.7	76.8	75.0
	ProLIP Fahes et al. (2026)	WACV'26	63.3	69.3	33.2	81.8	70.6	78.3	88.1	92.6	92.3	64.1	77.0	73.7
	Ours	—	**66.4**	**72.8**	**44.0**	**90.0**	**77.2**	**79.9**	**91.2**	**97.4**	**93.8**	**71.7**	**81.1**	**78.7**
16	TIP-Adapter-F Zhang et al. (2022)	ECCV'22	65.5	71.5	35.6	84.5	75.7	79.4	89.7	94.8	92.9	65.6	78.0	75.7
	TaskRes Yu et al. (2023)	CVPR'23	65.7	70.7	36.3	84.0	76.8	77.6	87.8	96.0	93.4	67.1	78.0	75.8
	GraphAdapter Li et al. (2023)	NeurIPS'23	65.7	71.2	36.9	85.3	76.2	78.6	88.6	96.2	93.3	67.6	78.8	76.2
	CLIP-Adapter Gao et al. (2024)	IJCV'24	63.5	69.5	32.1	84.4	74.0	78.2	87.8	93.9	92.4	65.9	76.7	74.9
	CLAP Silva-Rodriguez et al. (2024)	CVPR'24	65.0	70.8	33.6	80.1	75.1	78.5	88.5	94.2	91.9	66.4	76.3	74.6
	GDA Wang et al. (2024)	ICLR'24	63.8	70.7	40.6	86.1	75.1	79.0	88.8	95.7	92.5	66.5	77.5	76.0
	LP++ Huang et al. (2024)	CVPR'24	64.7	71.2	31.7	80.5	72.3	78.8	88.3	94.2	92.7	66.4	77.4	74.4
	CCA-FT Jiang et al. (2025)	ICCV'25	66.0	72.5	40.9	86.5	78.2	79.5	90.7	95.6	93.5	69.7	80.0	77.6
	ProKeR Bendou et al. (2025)	CVPR'25	64.5	71.8	40.2	80.8	76.1	79.4	89.0	96.1	93.5	67.6	78.3	76.8
	TIMO-S Li et al. (2025b)	AAAI'25	64.9	71.6	41.0	86.4	75.7	79.5	90.2	96.3	93.3	68.9	78.2	76.9
	LDC Li et al. (2025a)	CVPR'25	66.6	72.7	53.5	90.2	81.2	79.3	90.5	96.4	93.6	71.4	82.1	79.8
	SimNL Zhang et al. (2025)	WACV'25	66.5	72.3	40.2	87.3	77.4	79.8	90.5	96.5	93.7	70.8	80.2	77.8
	ProLIP Fahes et al. (2026)	WACV'26	64.3	71.3	39.9	85.8	75.6	79.2	89.1	94.9	93.4	68.0	79.7	76.5
	Ours	—	**67.5**	**73.4**	**58.3**	**92.2**	**82.0**	**80.4**	**92.0**	**98.1**	**94.0**	**73.4**	**83.1**	**81.3**

Table 11: **Full robustness analysis under matched support-set sampling.** Top-1 accuracy (%, mean ± std) over five independently sampled support sets using CLIP RN50. SAR and LDC use identical support sets and training protocols. The aggregate results are summarized in the main paper; here we report all dataset–shot combinations.

Shot	INet	SUN	Air	Euro	Cars	Food	Pets	Flow	Cal	DTD	UCF	Avg
SAR												
1	62.7±.04	66.0±.12	23.6±.20	79.6±.50	62.0±.08	78.1±.04	87.5±.43	87.3±.33	90.9±.12	57.8±.30	71.1±.21	69.7±.22
2	63.6±.12	68.9±.04	26.5±.10	84.5±.32	65.8±.23	78.5±.02	88.0±.11	92.7±.17	92.2±.22	62.5±.39	74.5±.34	72.5±.19
4	64.8±.10	70.9±.13	34.8±.87	89.3±.18	71.7±.14	78.9±.04	88.9±.30	94.8±.12	92.3±.18	66.3±.19	77.7±.22	75.5±.22
8	66.4±.02	72.6±.06	43.9±.37	89.8±.13	76.9±.17	79.8±.04	90.8±.07	97.2±.09	93.5±.22	71.4±.33	80.8±.31	78.5±.16
16	67.5±.04	73.6±.08	58.9±.48	92.1±.12	81.9±.09	80.3±.05	92.1±.12	98.2±.06	94.0±.13	73.2±.29	82.9±.21	81.3±.15
LDC												
1	60.5±.08	62.5±.17	19.1±.40	71.2±2.38	56.9±.23	77.0±.05	86.3±.48	80.4±.39	89.7±.39	51.1±1.09	67.2±.49	65.6±.56
2	61.3±.12	64.9±.10	21.7±.23	76.2±1.77	60.1±.11	77.3±.22	86.7±.54	86.2±.45	90.3±.23	55.8±1.37	70.7±.87	68.3±.55
4	62.5±.12	67.5±.07	28.6±.39	82.5±1.58	65.9±.63	77.9±.06	87.8±.40	91.2±.86	91.4±.23	61.8±1.22	74.3±.39	71.9±.54
8	64.3±.19	70.7±.12	39.9±.62	87.9±1.17	74.6±.32	78.7±.14	89.0±.34	94.0±.30	92.4±.20	66.7±.76	78.2±.55	76.0±.43
16	66.5±.01	72.8±.10	53.2±.38	90.9±.29	81.3±.14	79.4±.06	90.9±.17	96.4±.09	93.7±.11	71.3±.16	82.5±.29	79.9±.16

C.1 Complete Base-to-Novel Generalization

Table 12 complements the six benchmarks reported in Table 6 of the main paper with the remaining five datasets under the same ViT-B/16 MaPLe protocol. Together, the two tables cover all 11 standard benchmarks.

SAR achieves the strongest harmonic mean among the included protocol-matched baselines on each of the five additional datasets, extending the main-paper result to all 11 benchmarks. The improvement is not obtained by maximizing base-class accuracy alone. On OxfordFlowers, for example, SAR sacrifices 1.78 points relative to the strongest Base result but improves Novel accuracy by 1.72 points and HM by 3.26 points. Similar behavior appears on SUN397, where SAR is slightly below the strongest Base accuracy but improves Novel accuracy and achieves the best HM. This is

precisely the regime targeted by prior-preserving adaptation: task-specific discrimination improves without requiring the model to over-specialize to the observed base classes.

Across ImageNet, OxfordPets, and Food101, SAR improves both Novel accuracy and HM while retaining strong Base performance. Combined with the six datasets in the main paper, SAR therefore achieves the strongest HM among the included matched baselines on all 11 base-to-novel benchmarks. This result provides a complementary test of the paper's central preservation claim: selective intervention can improve the observed task while retaining semantic structure that remains useful for unseen classes.

Table 12: **Base-to-novel generalization on the five additional benchmarks.** Results use CLIP ViT-B/16 under the MaPLe protocol Khattak et al. (2023) and complement the six datasets reported in the main paper. We report Base and Novel accuracy (%) and their harmonic mean (HM). Δ_{best} is SAR's difference from the strongest included non-SAR result for the corresponding metric.

Dataset	Metric	CLIP	CoOp	CoCoOp	ProGrad	KgCoOp	MaPLe	Ours	Δ_{best}
ImageNet	Base	72.43	76.47	75.98	77.02	75.83	76.66	**77.93**	+0.91
	Novel	68.14	67.88	70.43	66.66	69.96	70.54	**71.26**	+0.72
	HM	70.22	71.92	73.10	71.46	72.78	73.47	**74.45**	+0.98
OxfordFlowers	Base	72.08	**97.60**	94.87	95.54	95.00	95.92	95.82	-1.78
	Novel	77.80	59.67	71.75	71.87	74.73	72.46	**79.52**	+1.72
	HM	74.83	74.06	81.71	82.03	83.65	82.56	**86.91**	+3.26
OxfordPets	Base	91.17	93.67	95.20	95.07	94.65	95.43	**95.62**	+0.19
	Novel	97.26	95.29	97.69	97.63	97.76	97.76	**98.34**	+0.58
	HM	94.12	94.47	96.43	96.33	96.18	96.58	**96.96**	+0.38
Food101	Base	90.10	88.33	90.70	90.37	90.50	90.71	**91.10**	+0.39
	Novel	91.22	82.26	91.29	89.59	91.70	92.05	**92.63**	+0.58
	HM	90.66	85.19	90.99	89.98	91.09	91.38	**91.86**	+0.48
SUN397	Base	69.36	80.60	79.74	81.26	80.29	**80.82**	80.60	-0.22
	Novel	75.35	65.89	76.86	74.17	76.53	78.70	**79.21**	+0.51
	HM	72.23	72.51	78.27	77.55	78.36	79.75	**79.90**	+0.15

C.2 Few-Shot Adaptation with ViT-B/16

The primary comparison in the main paper uses RN50 to preserve compatibility with the established adapter/cache benchmark. We separately evaluate SAR with CLIP ViT-B/16 to test whether its selective-repair principle transfers to a transformer visual hierarchy and to position it against recent ViT-based adaptation methods.

Table 13 reports the complete $1/2/4/8/16$-shot comparison across all 11 datasets. SAR achieves average accuracies of $76.2/78.7/80.7/82.4/84.4\%$, substantially improving its RN50 counterpart at every shot level. More importantly, the gains remain strong relative to recent ViT-B/16 approaches: SAR obtains the highest average accuracy among the included methods at 1-, 2-, and 4-shot, with margins of $+2.6$, $+2.0$, and $+0.2$ points over the strongest competing averages, respectively.

At 8 and 16 shots, MPS-Tuning becomes stronger in aggregate, while SAR remains competitive and continues to lead on several individual datasets. This trend is informative rather than contradictory to our motivation: the advantage of conservative, selective intervention is strongest when supervision is scarce, whereas methods with broader representation adaptation can benefit more as the support set grows. Crucially, the same SAR formulation transfers from RN50 residual stages to ViT transformer blocks without changing its central mechanism—class-complete evidence locates ambiguity, while the learned residual intervention remains restricted to the selected candidates.

Because recent ViT-B/16 results originate from several evaluation sources, Table 13 annotates values reproduced or reported through controlled evaluations in prior work. We therefore use this table to establish backbone transfer and contemporary positioning rather than to imply identical implementation pipelines across every method.

Table 13: **Few-shot comparison with CLIP ViT-B/16.** Top-1 accuracy (%) on 11 benchmarks over 1/2/4/8/16 shots. SAR achieves the strongest included cross-dataset average at 1, 2, and 4 shots and remains competitive at 8 and 16 shots. Values marked below the table originate from controlled ViT-B/16 evaluations or released results of the corresponding methods; otherwise values follow the cited papers. Best results within each dataset and shot setting are shown in **bold**.

	Method	Venue	INet	SUN	Air	Euro	Cars	Food	Pets	Flow	Cal	DTD	UCF	Avg.
1	CLIP-LoRA Zanella & Ben Ayed (2024)	CVPRW'24	**70.4**	70.4	30.2	72.3	70.1	84.3	**92.3**	83.2	93.7	54.3	76.3	72.5
	ITA Wang et al. (2025)	TIST'25	67.9	66.9	30.2	78.0	70.3	83.8	91.4	84.7	94.2	54.4	73.4	72.3
	MMA† Yang et al. (2024)	CVPR'24	69.2	64.0	28.7	55.1	67.9	83.0	91.2	83.6	92.9	52.3	74.2	69.3
	MMRL Guo & Gu (2025)	CVPR'25	69.0	68.9	28.5	76.0	68.7	84.9	90.9	86.0	94.2	56.4	76.0	72.7
	GalLoP‡ Lafon et al. (2024)	ECCV'24	69.8	68.8	**32.5**	63.6	71.5	84.8	91.6	86.1	94.1	57.2	73.1	72.1
	LDC§ Li et al. (2025a)	CVPR'25	69.5	68.0	27.6	78.5	68.2	85.9	91.3	83.6	94.2	58.3	73.3	72.6
	MPS-Tuning Chen et al. (2026)	ICLR'26	70.4	70.4	30.0	76.1	70.2	86.1	91.2	86.5	94.5	56.5	77.4	73.6
	TOGA Mohammad et al. (2026)	CVPR'26	69.2	68.1	31.0	67.4	69.1	86.2	91.2	88.2	94.3	55.2	74.5	72.2
	Ours	—	69.5	**71.5**	32.5	**84.8**	**72.4**	**86.7**	91.9	**91.2**	**95.0**	**64.6**	**78.3**	**76.2**
2	CLIP-LoRA Zanella & Ben Ayed (2024)	CVPRW'24	70.8	71.3	33.2	82.7	73.2	83.2	91.3	89.8	94.6	59.9	80.0	75.5
	ITA Wang et al. (2025)	TIST'25	70.0	72.0	32.8	82.7	74.6	85.8	92.7	90.8	94.8	59.4	78.6	75.8
	MMA† Yang et al. (2024)	CVPR'24	70.4	67.2	31.9	59.8	71.8	82.5	92.0	90.3	94.0	56.9	76.2	72.1
	MMRL Guo & Gu (2025)	CVPR'25	70.3	71.5	34.2	82.9	72.9	85.5	91.6	91.2	94.8	61.4	78.5	75.9
	GalLoP‡ Lafon et al. (2024)	ECCV'24	70.6	71.2	**36.7**	67.8	75.0	85.4	**93.1**	91.8	95.3	61.9	77.8	75.2
	LDC§ Li et al. (2025a)	CVPR'25	69.9	69.6	29.0	81.7	70.8	86.1	91.2	89.2	94.6	62.2	76.0	74.6
	MPS-Tuning Chen et al. (2026)	ICLR'26	**71.4**	72.4	33.8	81.6	75.4	86.7	92.4	91.6	95.5	61.5	81.0	76.7
	TOGA Mohammad et al. (2026)	CVPR'26	69.9	70.7	34.8	74.9	72.8	86.5	92.2	91.9	95.0	58.4	77.9	75.0
	SurPL* Liu et al. (2025)	ICML'25	–	–	–	–	–	–	–	–	–	–	–	75.4
	Ours	—	71.0	**74.1**	36.2	**87.9**	**76.1**	**86.9**	92.2	**95.7**	**95.8**	**68.5**	**81.7**	**78.7**
4	CLIP-LoRA Zanella & Ben Ayed (2024)	CVPRW'24	71.4	72.8	37.9	84.9	77.4	82.7	91.0	93.7	95.2	63.8	81.1	77.4
	ITA Wang et al. (2025)	TIST'25	71.3	74.0	38.2	86.2	77.1	86.2	93.6	93.5	95.6	66.9	80.9	78.5
	MMA† Yang et al. (2024)	CVPR'24	71.0	70.0	37.6	79.4	76.5	82.1	92.2	93.0	94.3	63.9	80.1	76.4
	MMRL Guo & Gu (2025)	CVPR'25	71.4	73.9	40.5	87.7	78.2	85.8	92.6	94.6	96.0	67.9	82.7	79.2
	GalLoP‡ Lafon et al. (2024)	ECCV'24	71.7	73.4	**42.7**	74.1	79.7	86.1	93.2	95.6	95.6	66.9	81.2	78.2
	LDC§ Li et al. (2025a)	CVPR'25	71.0	73.1	32.0	86.4	75.1	86.7	91.8	94.1	95.3	66.4	79.8	77.4
	MPS-Tuning Chen et al. (2026)	ICLR'26	**72.6**	74.5	42.0	88.5	**81.2**	87.2	**93.7**	96.0	**96.8**	68.5	**84.3**	80.5
	TOGA Mohammad et al. (2026)	CVPR'26	71.0	73.2	38.3	80.3	76.7	86.7	92.7	96.4	95.7	64.5	81.7	77.9
	SurPL* Liu et al. (2025)	ICML'25	–	–	–	–	–	–	–	–	–	–	–	79.2
	Ours	—	72.3	**76.1**	40.3	**90.1**	78.2	**87.4**	93.3	**97.5**	96.1	**73.8**	82.8	**80.7**
8	CLIP-LoRA Zanella & Ben Ayed (2024)	CVPRW'24	72.3	74.7	45.7	89.7	82.1	83.1	91.7	96.3	95.6	67.5	84.1	80.3
	ITA Wang et al. (2025)	TIST'25	72.4	75.2	42.0	89.2	79.2	85.5	**93.8**	95.8	95.8	68.9	84.7	80.2
	MMA† Yang et al. (2024)	CVPR'24	71.8	72.3	44.8	86.5	81.4	83.0	92.8	96.0	95.4	68.0	83.4	79.6
	MMRL Guo & Gu (2025)	CVPR'25	72.3	76.0	48.1	88.7	82.6	86.3	93.0	96.6	96.3	71.6	84.7	81.5
	GalLoP‡ Lafon et al. (2024)	ECCV'24	72.9	75.5	48.5	84.0	84.4	86.8	93.7	97.9	96.2	71.9	84.4	81.5
	LDC§ Li et al. (2025a)	CVPR'25	72.5	75.6	38.0	90.8	79.7	86.9	92.5	96.0	95.9	71.6	81.3	80.1
	MPS-Tuning Chen et al. (2026)	ICLR'26	**74.1**	76.7	**55.6**	91.1	**87.2**	**87.6**	93.8	98.0	**96.8**	73.3	**87.0**	**83.8**
	TOGA Mohammad et al. (2026)	CVPR'26	71.5	75.1	44.2	84.1	78.2	86.8	93.4	97.3	95.8	69.6	83.9	80.0
	SurPL* Liu et al. (2025)	ICML'25	–	–	–	–	–	–	–	–	–	–	–	82.5
	Ours	—	72.3	**77.3**	43.8	**92.8**	82.4	87.6	**93.8**	**98.6**	96.3	**75.5**	84.5	82.4
16	CLIP-LoRA Zanella & Ben Ayed (2024)	CVPRW'24	73.6	76.1	54.7	92.1	86.3	84.2	92.4	98.0	96.4	72.0	86.7	83.0
	ITA Wang et al. (2025)	TIST'25	73.0	77.0	47.9	92.9	82.7	87.7	93.9	97.5	96.1	72.8	85.8	82.5
	MMA† Yang et al. (2024)	CVPR'24	73.1	74.6	52.7	92.4	85.7	84.6	93.2	98.0	96.3	73.5	86.3	82.8
	MMRL Guo & Gu (2025)	CVPR'25	73.4	77.7	57.6	93.4	86.4	87.0	93.8	98.4	97.1	75.3	87.6	84.3
	GalLoP‡ Lafon et al. (2024)	ECCV'24	74.1	77.3	55.2	89.2	88.2	87.4	94.3	98.7	96.6	75.0	86.6	83.9
	LDC§ Li et al. (2025a)	CVPR'25	73.9	76.9	47.9	92.2	84.2	87.4	93.4	97.9	96.6	77.3	84.8	82.9
	MPS-Tuning Chen et al. (2026)	ICLR'26	**75.6**	78.5	**69.0**	**94.4**	**91.1**	**88.1**	94.8	**99.4**	**97.4**	77.2	**89.9**	**86.9**
	TOGA Mohammad et al. (2026)	CVPR'26	72.3	76.2	48.4	89.4	85.3	87.5	93.4	98.3	96.3	73.6	84.9	82.3
	SurPL* Liu et al. (2025)	ICML'25	74.7	77.7	60.5	93.9	89.0	87.6	94.2	98.8	96.9	74.8	88.2	85.1
	Ours	—	74.0	**78.6**	53.4	93.8	85.5	87.8	**94.8**	98.8	96.6	**78.4**	86.3	84.4

† MMA multi-shot values are from the controlled ViT-B/16 evaluation reported by MMRL. ‡ GalLoP multi-shot values are from the controlled ViT-B/16 evaluation reported by MPS-Tuning. § LDC values are from the authors' released ViT-B/16 results. * SurPL reports dataset-wise ViT-B/16 values for the 16-shot setting, while only overall averages are available for the 2/4/8-shot settings.

D Mechanism-Level Analysis

The main paper establishes three complementary properties of SAR: it reduces ambiguity among competing candidates, concentrates prediction changes within a small semantic neighborhood, and retains a recovery path when candidate selection is incomplete. This section expands those analyses with per-shot diagnostics and additional locality and recovery results.

D.1 Formal Property of Selective Repair

Proposition 1 (Localized bounded intervention)**.** *Let $z^{(0)} \in \mathbb{R}^C$ be the CLIP-anchored proposal and $\mathcal{I}_k$ the selected candidate set. Under Eqs. 6–8, the selectively repaired logits satisfy*

$$z_j^{\mathrm{sel}} = z_j^{(0)}, \qquad \forall j \notin \mathcal{I}_k,$$

and

$$\left| z_j^{\mathrm{sel}} - z_j^{(0)} \right| \le \eta, \qquad \forall j \in \mathcal{I}_k.$$

Consequently,

$$\mathrm{supp}\Big(z^{\mathrm{sel}} - z^{(0)}\Big) \subseteq \mathcal{I}_k, \qquad \Big\| z^{\mathrm{sel}} - z^{(0)} \Big\|_\infty \le \eta.$$

Proof. By construction, the residual r^{sel} is zero outside $\mathcal{I}_k$. For selected classes, $r_j^{\mathrm{sel}} = \delta_t$, where $\delta = \eta \tanh(\cdot)$; hence $|\delta_t| \le \eta$. Since $z^{\mathrm{sel}} = z^{(0)} + r^{\mathrm{sel}}$, both statements follow directly. □

Corollary 1 (Exact preservation outside the candidate set)**.** *For any $i, j \notin \mathcal{I}_k$,*

$$z_i^{\mathrm{sel}} - z_j^{\mathrm{sel}} = z_i^{(0)} - z_j^{(0)}.$$

Thus, selective repair exactly preserves the pairwise ordering and margins among all non-candidate classes.

Proposition 1 formalizes what is selective in SAR: the repair branch can alter at most k logits and each direct intervention is bounded by η. Corollary 1 further shows that the selective branch cannot disturb relative evidence among classes outside the ambiguity set, providing a formal counterpart to the empirical repair–preserve and locality analyses.

D.2 Per-Shot Ambiguity Resolution

Table 14 expands the aggregate ambiguity diagnostics reported in Table 3 of the main paper. The evaluation set is fixed *before adaptation*: we retain samples misclassified by frozen CLIP for which the ground-truth class lies inside CLIP's original top-k candidate set. We then measure normalized candidate entropy H_k, signed ground-truth margin $M_y = z_y - \max_{c \neq y} z_c$, and the fraction of initially negative margins that become positive after adaptation (Resolved).

The per-shot results reinforce the aggregate finding. Top-k SAR achieves the lowest candidate entropy and the highest ambiguity-resolution rate at every shot level. Its signed ground-truth margin is also strongest overall, although individual competing variants can be marginally higher at a specific shot. The comparison with Random-k is particularly informative: both methods modify the same number of classes, yet targeted repair consistently resolves more ambiguity. The gain therefore comes from *where* the residual is applied rather than sparsity alone.

Calibration. We additionally report ECE and NLL on the complete test set in Table 15. Adaptation substantially improves NLL over frozen CLIP as additional supervision becomes available. Among the adapted variants, Top-k SAR is generally best or near-best, although the calibration advantage is not uniform at every shot. Notably, frozen CLIP retains lower ECE than all adapted models in the lowest-shot regimes, highlighting that increased task discrimination and calibration are distinct objectives.

D.3 Locality of the Learned Intervention

Fig. 5 provides a complementary view of the repair–preserve analysis in Fig. 3 of the main paper. For each frozen-CLIP error, we rank classes by the magnitude of their logit change and measure how much of the total absolute change is captured by the r most-modified classes.

Across 1/2/4-shot settings, SAR concentrates a substantially larger fraction of its total score change in the first few classes than dense correction. This distinction is important: SAR does not avoid class-complete evidence computation, but its *learned intervention* is localized once ambiguity has been identified. The result therefore provides direct behavioral evidence for the defining property in Eq. 8: all classes can contribute to locating ambiguity, while only a small candidate set receives the selective residual.

Table 14: **Per-shot ambiguity-resolution diagnostics.** Metrics are computed on the fixed subset of frozen-CLIP errors whose ground-truth class lies within CLIP's original top-k candidate set. Lower H_k and higher M_y/Resolved are better.

Shot	Method	$H_k \downarrow$	$M_y \uparrow$	Resolved (%) ↑
1	Frozen CLIP	.761	-1.044	0.000
	No repair	.725	-.002	45.015
	Random-k	.721	-.002	45.237
	Dense repair	.725	.008	45.038
	Top-k SAR	**.713**	**.025**	**45.875**
2	Frozen CLIP	.816	-.866	0.000
	No repair	.682	.439	55.533
	Random-k	.683	.433	55.004
	Dense repair	.677	.377	54.351
	Top-k SAR	**.671**	**.443**	**56.550**
4	Frozen CLIP	.788	-.974	0.000
	No repair	.606	.600	58.398
	Random-k	.607	.614	59.367
	Dense repair	.608	.566	59.432
	Top-k SAR	**.596**	**.625**	**59.693**
8	Frozen CLIP	.799	-.912	0.000
	No repair	.539	1.027	66.745
	Random-k	.536	**1.030**	66.534
	Dense repair	.544	1.022	66.445
	Top-k SAR	**.530**	1.028	**66.937**
16	Frozen CLIP	.818	-.928	0.000
	No repair	.469	1.522	70.659
	Random-k	.468	1.534	71.120
	Dense repair	.462	1.537	71.319
	Top-k SAR	**.454**	**1.542**	**71.442**

Table 15: **Calibration diagnostics on the complete test set.** ECE and NLL are reported for each shot setting; lower is better.

	1-shot		2-shot		4-shot		8-shot		16-shot	
Method	ECE	NLL	ECE	NLL	ECE	NLL	ECE	NLL	ECE	NLL
Frozen CLIP	.042	1.568	.042	1.568	.042	1.568	.042	1.568	.042	1.568
No repair	.237	1.470	.203	1.262	.172	1.075	.141	.927	.082	.749
Random-k	.224	1.462	.207	1.266	.171	1.079	.131	**.911**	.084	.749
Dense	.226	1.477	**.196**	1.249	.179	1.064	.137	.927	.087	.757
Top-k SAR	**.215**	**1.400**	.198	**1.241**	**.155**	**1.057**	**.125**	.912	**.077**	**.740**

D.4 Recovery Beyond the Selected Candidate Set

Table 4 in the main paper shows that SAR remains effective when candidate selection does not contain the ground truth. Here we provide additional analysis of that failure regime.

The experiment pools initial frozen-CLIP errors from seven datasets (EuroSAT, Caltech101, FGVC-Aircraft, DTD, Food101, StanfordCars, and SUN397) at 1/4/16 shots using RN50. Across the resulting 21 dataset–shot configurations, 36.4% of errors place the ground truth outside the selected ambiguity set, making this a substantial rather than exceptional case. The final SAR prediction improves over LDC in all three pooled rank strata: $+4.75$ points when the GT lies inside the selected set, $+6.94$ points when it lies within CLIP's top-10 but outside the selected set, and $+4.72$ points when it lies outside the top-10.

The repaired branch alone does not explain this robustness. When the GT lies outside the selected set, z^{sel} cannot directly increase its score through the selective residual. The improvement appears only after fusion with the class-complete z^{full} branch, validating its intended role as a recovery path. Across the nine shot-by-rank strata, final SAR outperforms LDC in eight; the only exception occurs at 16-shot for GTs within the top-10 but outside the selected set, where LDC obtains 26.07% recovery versus 24.28% for SAR. When this critical subgroup is pooled by dataset across shots, however, SAR improves over LDC on all seven datasets.



Figure 5: **Selective repair produces substantially more localized logit updates.** For frozen-CLIP errors, classes are ranked by the magnitude of their logit change, and each panel reports the cumulative fraction of total absolute change captured by the r most-modified classes at 1/2/4 shots (RN50, aggregated over 10 downstream datasets; shaded regions denote 95% bootstrap confidence intervals). The selective repair branch concentrates markedly more of its update in a small number of classes than dense full-class correction, consistently across shot levels. This provides direct behavioral evidence for SAR's central design: class-complete evidence is used to locate ambiguity, while the learned repair itself remains localized to ambiguity-relevant candidates.

Recovery naturally decreases as CLIP ranks the true class farther from its initial semantic neighborhood. This exposes a meaningful failure boundary: the full-class branch mitigates imperfect candidate selection, but it cannot fully compensate when the pretrained prior itself is severely misaligned.

E Additional Ablation Studies

The main paper isolates SAR's two central design choices: *where* repair should be applied and *which depth evidence* should support it (Table 5). Here we analyze the remaining design decisions: the size and inputs of the ambiguity set, the training objective, and the role of shot-adaptive proposal and prediction fusion. Unless stated otherwise, results are averaged over the 11 RN50 few-shot benchmarks at 1/4/16 shots.

E.1 Candidate Set and Repair Evidence

How large should the ambiguity set be? Table 16 varies the repair budget k on ImageNet, FGVC-Aircraft, EuroSAT, and Food101. Performance does not improve monotonically as more classes are made editable. Instead, a compact candidate set performs best, with $k = 5$ giving the highest average accuracy at 1-, 4-, and 16-shot in this sweep. Expanding the intervention to ten classes already weakens performance, while dense repair over all C classes is consistently worst.

This behavior supports the intended role of candidate selection. The top-k set is not simply a computational approximation to full-class correction; it acts as an inductive constraint on *where adaptation is allowed to intervene*. Once the set becomes too broad, the repair increasingly resembles dense correction and loses the preservation advantage demonstrated in the main-paper repair–preserve analysis.

Table 16: **Effect of candidate-set size.** Average top-1 accuracy (%) on ImageNet, FGVC-Aircraft, EuroSAT, and Food101 as the repair budget k varies. A compact ambiguity set performs best, while dense intervention ($k = C$) is consistently weaker.

k	1	2	3	5	10	C (Dense)
1-shot	59.0	59.4	59.7	**60.9**	60.0	58.4
4-shot	64.3	65.6	65.7	**67.0**	66.2	63.9
16-shot	72.6	72.5	73.0	**74.6**	73.5	71.7

What information does the repair head need? Table 17 progressively augments the information available to the selective repair head. The original CLIP candidate scores o already provide a useful

Table 17: **Repair-input ablation.** Average top-1 accuracy (%) as the repair head progressively receives the original CLIP candidate scores o, proposal discrepancy d, ambiguity margin ξ_k, and fused visual evidence $\bar{f}$.

Shot	o	$o+d$	$o+d+\xi_k$	$\bar{f}+o+d+\xi_k$
1-shot	68.4	68.7	68.8	**69.7**
4-shot	74.5	74.9	75.1	**75.6**
16-shot	80.2	80.5	81.0	**81.3**

Table 18: **Training-objective ablation.** Average top-1 accuracy (%) after removing individual supervision or prior-anchoring terms from Eq. 10.

Shot	Full obj.	$-\mathcal{L}_{\mathrm{CE}}^{\mathrm{full}}$	$-\mathcal{L}_{\mathrm{CE}}^{\mathrm{sel}}$	$-\mathcal{L}_{\mathrm{CE}}^{\mathrm{final}}$	$-\mathcal{L}_{1}^{\mathrm{full}}$	$-\mathcal{L}_{1}^{\mathrm{sel}}$	$-\mathcal{L}_{1}^{\mathrm{both}}$
1-shot	**69.7**	50.3	59.6	58.1	45.8	47.8	52.8
4-shot	**75.6**	62.5	69.6	69.6	64.4	65.5	68.1
16-shot	**81.3**	72.4	77.4	79.2	77.8	78.4	79.3
Average	**75.5**	61.7	68.9	69.0	62.7	63.9	66.7

semantic prior over the competing classes. Adding the proposal discrepancy d consistently improves performance, indicating that repair benefits from knowing how the adapted proposal departs from the frozen prior. The ambiguity margin ξ_k provides additional information about the separation between the leading candidates.

The largest improvement comes from additionally supplying the fused visual representation $\bar{f}$. This distinction is central to SAR: the candidate scores indicate *where* ambiguity lies, whereas multi-depth visual evidence helps determine *how* those candidates should be separated. Their combination yields the strongest result at every shot level.

E.2 Training Objective

Table 18 tests the contribution of each term in Eq. 10. The three cross-entropy terms supervise the full-class recovery branch z^{full}, the selectively repaired branch z^{sel}, and the fused prediction z, while the two ℓ_1 terms anchor the branch predictions to the frozen CLIP prior.

Direct supervision of z^{full} is especially important: removing its cross-entropy term produces the largest average degradation. This is consistent with the role of the recovery branch; multi-depth visual evidence must remain independently discriminative rather than relying on the final gate to compensate for a weak class-complete prediction. Supervising z^{sel} and the final prediction is likewise necessary, showing that effective fusion does not emerge from branch-level learning alone.

The prior-anchoring terms are also consequential, particularly in the low-shot regime. Removing either one permits substantially greater deviation from the frozen prediction, while removing both weakens the repair–preserve balance further. The full objective therefore couples two requirements: each branch must become discriminative under task supervision, but neither should be free to overwrite the pretrained decision structure without constraint.

E.3 Shot Adaptivity and Prediction Fusion

SAR uses the amount of available supervision in two places: when constructing the CLIP-anchored proposal and when combining the selective and full-class branches. Tables 19 and 20 isolate these effects.

Where is shot conditioning useful? Table 19 compares fixed-capacity alternatives with shot-adaptive proposal and fusion. Neither a uniformly shallow nor uniformly deep transformation is sufficient: the former is overly restrictive as supervision increases, whereas the latter introduces unnecessary capacity in the lowest-shot regime. Conditioning either the proposal or the final gate

on K improves over these fixed alternatives, but applying shot adaptivity to both is consistently strongest.

The result supports a simple principle: the amount of learned deviation from the pretrained prior should track the reliability of the available support. With very few examples, SAR remains conservative; as supervision increases, it can rely more strongly on task-specific evidence without abandoning the pretrained reference.

Table 19: **Ablation of shot-adaptive components.** Average top-1 accuracy (%) when proposal capacity and final prediction fusion are fixed or conditioned on the number of shots.

Method	1-shot	4-shot	16-shot
Fixed Shallow	62.1	70.1	78.3
Fixed Deep	62.6	70.4	78.9
Shot-adaptive proposal Only	63.2	70.9	79.2
Shot-adaptive gate Only	64.5	71.3	79.5
Full shot-adaptive	**69.7**	**75.6**	**81.3**

Why retain both prediction branches? Table 20 compares the full-class recovery branch z^{full}, the selectively repaired branch z^{sel}, and several fusion strategies. Neither branch alone matches their combination. The full-class branch is substantially stronger in isolation, particularly as the number of shots increases, reflecting the discriminative value of learned multi-depth evidence. The selective branch is intentionally more conservative: it remains anchored to CLIP and restricts direct repair to the ambiguity set.

Simple averaging already improves over either branch alone, confirming their complementarity. A sample-adaptive gate without shot conditioning provides similar performance, but the full shot-adaptive gate is consistently stronger, with the gap increasing from 0.8 points at 1-shot to 1.5 points at 16-shot relative to the non-shot-conditioned gate. Thus, fusion is not merely an ensemble effect; SAR benefits from explicitly adjusting the balance between prior-preserving local repair and class-complete task-specific evidence as supervision changes.

F Computational Analysis

SAR is selective in *where the learned prediction correction is applied*; it does not claim sparse computation throughout the adaptation pipeline. Depth-relational evidence extraction and the full-class recovery branch remain class-complete by design. We therefore quantify the computational consequence of this choice through a matched comparison with LDC, the closest dense logit-correction baseline used throughout the paper.

F.1 Matched Computational Comparison

Table 21 compares SAR and the official LDC implementation under the same ImageNet 16-shot RN50 configuration on a single NVIDIA A40. Both methods use 224×224 inputs and FP16 model execution. GFLOPs are measured for one complete forward pass per image; memory is reported at batch size 64 using both peak allocated and peak reserved CUDA memory.

SAR has 124.64M total parameters, of which 22.63M are trainable, compared with 119.26M and 17.26M for LDC. Thus, SAR introduces 4.5% more total parameters and 31.2% more trainable parameters. Its forward computation increases from 19.86 to 24.27 GFLOPs per image ($+22.2\%$), reflecting the additional multi-depth evidence extraction and routed repair components.

The increase in arithmetic does not translate into a comparable memory increase. At batch size 64, peak allocated training memory changes only from 3.85 to 3.89 GB ($+1.1\%$), while peak reserved memory increases from 5.39 to 5.66 GB ($+5.0\%$). At inference, allocated memory changes from 2.20 to 2.22 GB ($+0.5\%$). Thus, SAR trades moderate additional computation and trainable capacity for substantially more targeted prediction repair, while leaving the practical GPU-memory footprint nearly unchanged.

Table 20: **Prediction-branch fusion ablation.** Average top-1 accuracy (%) for the full-class branch z^{full}, selective branch z^{sel}, and alternative fusion strategies.

Shot	z^{full} only	z^{sel} only	Fixed $\omega = .5$	Adaptive, no shot	**Full SAR**
1-shot	66.3	64.0	68.9	68.9	**69.7**
4-shot	73.7	69.3	74.4	74.3	**75.6**
16-shot	79.4	72.9	79.7	79.8	**81.3**

Table 21: **Matched computational comparison with LDC.** Measurements use CLIP RN50 on ImageNet 16-shot with 224×224 inputs and FP16 execution on the same NVIDIA A40. Training and inference memory are reported at batch size 64 as peak allocated / peak reserved CUDA memory. GFLOPs denote one complete forward pass per image.

Method	Total params	Trainable params	GFLOPs / image	Train VRAM (GB)	Inference VRAM (GB)
LDC	119.26M	17.26M	19.86	3.85 / 5.39	2.20 / 2.49
SAR	124.64M	22.63M	24.27	3.89 / 5.66	2.22 / 2.50
Relative change	+4.51%	+31.17%	+22.24%	+1.12% / +4.98%	+0.49% / +0.42%

F.2 Measurement Details

For consistency, all measurements in Table 21 are obtained on the same hardware and configuration after warm-up. Peak allocated memory is measured with `torch.cuda.max_memory_allocated()`, while peak reserved memory uses `torch.cuda.max_memory_reserved()`. GFLOPs are measured with `torch.profiler(with_flops=True)`, where a multiply–add contributes two floating-point operations. Text features are precomputed for both methods, so the comparison reflects the visual backbone and active adaptation modules rather than repeated text encoding.

We report both allocated and reserved memory because they measure different quantities: allocated memory reflects live tensors, whereas reserved memory also includes memory retained by PyTorch's CUDA allocator. This distinction explains why process-level GPU-memory readings can differ substantially from peak tensor allocation and avoids conflating memory definitions across implementations.

G Qualitative Analysis

The quantitative analyses in the main paper establish that SAR improves ambiguity resolution, concentrates prediction changes within a small semantic neighborhood, and better preserves initially reliable CLIP decisions. Here we complement those aggregate results with sample-level examples. The three qualitative analyses serve different purposes: the first illustrates the repair–preserve behavior summarized in the main paper; the second exposes distinct operating regimes, including a failure boundary; and the third focuses specifically on fine-grained near-neighbor confusions. These examples are intended to illustrate the observed behaviors rather than to estimate their frequency, which is quantified by the corresponding aggregate analyses in the main paper.

G.1 Repair–Preserve Behavior

Figure 6 provides the qualitative counterpart to the repair–preserve analysis in the main paper. The examples show that SAR's advantage is not simply a consequence of making larger corrections.

In panels A and D, frozen CLIP and dense logit correction (LDC) retain an incorrect prediction, whereas SAR promotes the ground-truth class to top-1. Panel B represents a different regime: both LDC and SAR repair the original CLIP error, but SAR produces a substantially larger ground-truth margin, indicating more decisive resolution of the local ambiguity. Panel C illustrates the preservation side of the objective. Frozen CLIP is already correct but uncertain; dense correction changes the decision, whereas SAR preserves the correct class and increases its confidence.

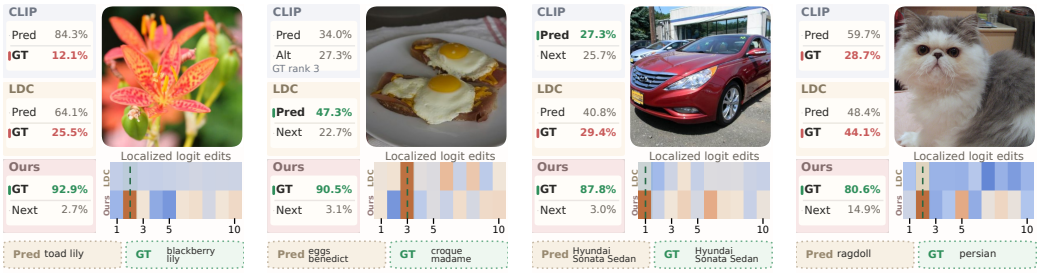

Figure 6: **Qualitative repair–preserve behavior of SAR (left-to-right). A, D:** Frozen CLIP and dense logit correction (LDC) remain incorrect, whereas SAR recovers the ground truth. **B:** Both adaptation methods correct the CLIP error, but SAR produces a substantially larger ground-truth margin. **C:** CLIP is correct but uncertain; LDC flips the decision, whereas SAR preserves and strengthens the correct prediction. The examples illustrate that selective repair can simultaneously correct local ambiguity and reduce unnecessary disruption of reliable pretrained decisions.

Taken together, these examples illustrate the two-sided behavior measured quantitatively in the main paper: useful adaptation must not only repair incorrect predictions, but also avoid unnecessarily disturbing decisions for which the pretrained prior remains reliable.

G.2 Operating Regimes and Failure Boundary

Figure 7 examines a broader range of behaviors, including cases in which selective intervention is beneficial but not sufficient.

Bounded but imperfect correction. In the Caltech101 example, frozen CLIP initially places substantial probability on the ground truth, while dense correction shifts strongly toward an incorrect alternative. SAR also fails to retain the correct top-1 decision, but keeps the ground-truth score much closer to the competing prediction. This case illustrates a useful consequence of bounded intervention: even when repair is incomplete, the model can avoid the severe drift induced by a full-distribution correction.

Preserving a reliable pretrained decision. In the Stanford Cars example, CLIP is correct but uncertain. LDC changes the top-1 prediction, whereas SAR preserves the ground truth and increases its separation from the nearest competitor. This provides a sample-level illustration of the preservation behavior measured by the repair–preserve analysis.

Resolving local ambiguity. For Oxford Flowers, the ground-truth class begins at rank 4 under frozen CLIP, so the error lies within a plausible local semantic neighborhood. Both adapted methods recover the class, but SAR assigns it a substantially larger margin over the remaining candidates. This is the regime for which query-specific selective repair is explicitly designed.

Failure when the pretrained neighborhood is poor. The SUN397 example exposes the complementary limitation. The ground-truth class begins at rank 83 under frozen CLIP, far outside the local ambiguity region. LDC recovers it by modifying the full class distribution, whereas SAR remains more conservative and moves the class only to rank 5. This example clarifies the trade-off introduced by selective intervention: when the pretrained ranking is severely misaligned, locality can become a constraint rather than an advantage. The full-class recovery branch mitigates this failure mode at the population level, as shown by the outside-top-k analysis, but it does not eliminate it.

G.3 Fine-Grained Near-Neighbor Confusions

The previous figures span different repair and preservation regimes. Figure 8 instead isolates a narrower but important failure mode: confusion between semantically or visually similar classes.

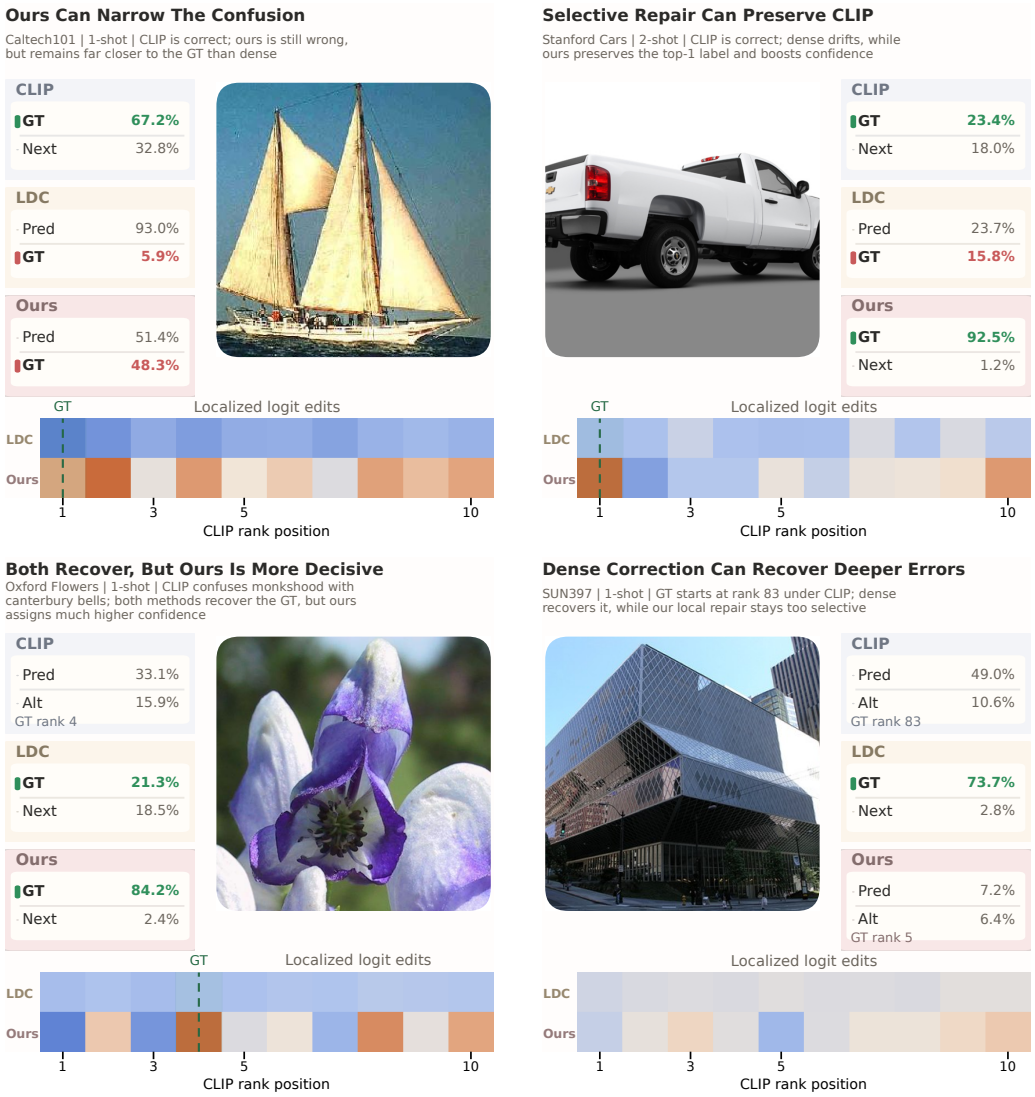

Figure 7: **Operating regimes and failure modes of selective ambiguity repair.** We compare frozen CLIP, dense logit correction (LDC), and SAR across four representative cases. **Top-left: Bounded but imperfect correction.** Dense adaptation strongly moves away from a reliable CLIP prediction, whereas SAR remains much closer to the ground truth but does not fully recover top-1. **Top-right: Prior preservation.** CLIP is correct but uncertain; LDC flips the decision, while SAR preserves and strengthens it. **Bottom-left: Local ambiguity repair.** The ground truth begins at rank 4 and is recovered by both methods, with SAR producing a larger margin. **Bottom-right: Failure boundary.** The ground truth begins at rank 83; dense correction recovers it, whereas SAR remains conservative and moves it only to rank 5. Score panels report prediction confidence and ground-truth rank where applicable, while heatmaps visualize the corresponding logit changes over CLIP-ranked candidates.

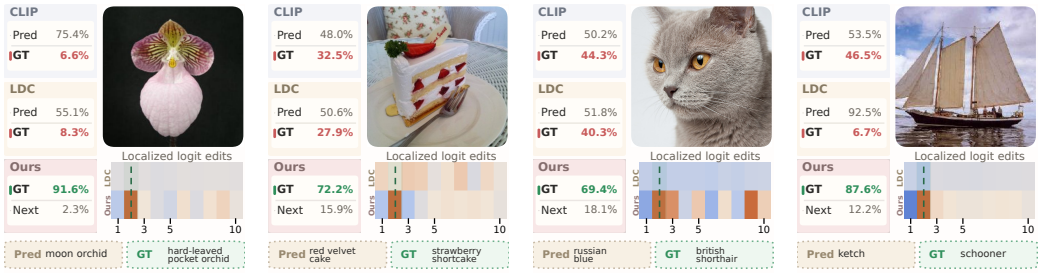

Figure 8: **Fine-grained near-neighbor ambiguity repair.** Each example shows a query for which frozen CLIP assigns higher confidence to a visually or semantically related class than to the ground truth. Dense logit correction (LDC) partially changes the competing scores but retains the incorrect ordering, whereas SAR promotes the ground truth above the confuser. Heatmaps visualize the magnitude and positions of logit changes over CLIP-ranked candidates; the dominant SAR updates remain concentrated around the high-ranked ambiguity neighborhood rather than being distributed broadly across classes.

Across the four examples, frozen CLIP places the ground truth within a plausible high-ranked neighborhood but favors a closely related alternative: *moon orchid* over *hard-leaved pocket orchid*, *red velvet cake* over *strawberry shortcake*, *Russian blue* over *British shorthair*, and *ketch* over *schooner*. LDC modifies the competing scores but retains the incorrect ordering, whereas SAR promotes the ground truth above the confuser in each case.

The corresponding logit-change patterns are consistent with the intended mechanism: the dominant modifications remain concentrated around the high-ranked candidate neighborhood rather than being distributed broadly over the class space.

These examples therefore complement the aggregate locality analysis in Appendix D.3: the benefit of selective repair is not that it changes logits more aggressively, but that it places the intervention on the classes already implicated by the query-specific ambiguity.