REPURPOSE: FROM PREDICTION TO CORRECTION WITH TASK-SPECIALIZED REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Base-to-novel adaptation of CLIP aims to learn task-specific knowledge from base classes while preserving generalization to unseen classes. Prior work has shown that adaptation-specific representations learned through inserted tokens can become highly task-specialized and perform poorly when directly used for novel-class prediction, limiting their role in novel-class inference. We revisit these representations from a different perspective and introduce REPURPOSE, a training-free inference framework that repurposes REP from prediction to correction: instead of asking REP to predict novel classes directly, REPURPOSE uses its sample-specific information to propose constrained corrections to CLS decisions and verifies them with frozen CLIP semantics. Across 11 base-to-novel benchmarks, applying REPURPOSE to MMRL++ improves its average Novel accuracy from 78.32% to 79.25%, achieving state-of-the-art performance among methods without knowledge distillation. Beyond MMRL++, the same correction principle consistently improves multiple adaptation hosts and transfer settings, and further extends to the autoregressive M2PT inference pipeline without modifying decoding.

1 INTRODUCTION

Vision-language models (VLMs) have demonstrated strong transferability by learning aligned representations across visual and textual modalities. Among them, CLIP (Radford et al., 2021), pretrained on large-scale image-text pairs, exhibits remarkable zero-shot recognition ability and has become a widely used foundation model for downstream vision tasks (Zhou et al., 2022b). To efficiently adapt pretrained VLMs to specific downstream tasks, prompt learning methods introduce learnable textual prompts, visual prompts, or cross-modal interactions while largely preserving the pretrained model (Zhou et al., 2022b;a; Khattak et al., 2023a).

Despite their effectiveness, these methods are typically optimized with supervision from base classes, which can shift the adapted model toward task-specific information that does not transfer well to unseen categories (Zhou et al., 2022a). This tension between downstream adaptation and generalization has become a central challenge in base-to-novel recognition. Recent methods have sought to mitigate this issue by preserving transferable knowledge during adaptation. PromptSRC regularizes the adapted model toward its pretrained counterpart, while DePT disentangles base-specific and task-shared representations (Khattak et al., 2023b; Zhang et al., 2024).

Another line of work explicitly captures adaptation-specific information through dedicated representations or representation pathways. MMRL and MMRL++ introduce additional representation tokens (REP) to provide a dedicated pathway for learning task- and dataset-related information while maintaining CLS as the primary transferable representation (Guo & Gu, 2025; 2026). As illustrated in Fig. 1 (left), during base-class inference, REP contributes directly to prediction by computing similarity with text embeddings alongside CLS, and the resulting scores are combined for classification. This design allows REP to capture adaptation-specific information that complements the pretrained semantic knowledge retained in CLS. The role of REP, however, is different for novel-class inference. MMRL adopts a decoupled inference strategy that uses both CLS and REP for base classes but relies only on CLS for unseen classes. Its ablation study provides a clear motivation for this design: when REP is also incorporated into novel-class prediction (w/o DS2), novel accuracy drops from 77.16% to 73.80%, a decrease of 3.36 percentage points (Guo & Gu, 2025). This result suggests that directly transferring the base-class use of REP to novel classes can be detrimental. More

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

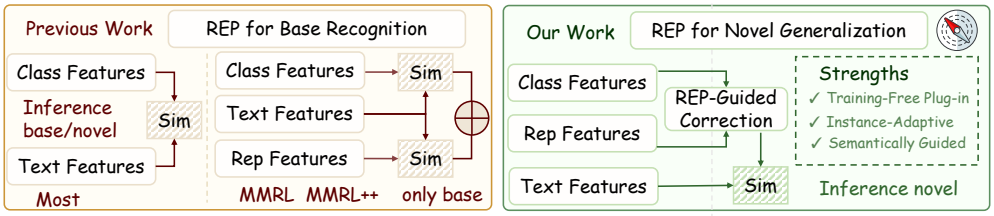

Figure 1: **Motivation of REPURPOSE.** Adapted hosts produce specialized states beyond the final CLS prediction representation. We read such states as REP and repurpose them as constrained corrective evidence for CLS rather than as standalone predictors.

broadly, prior work has also observed that specialized states induced by inserted tokens may be less effective than the conventional CLS representation when directly used for prediction (Chowdhury et al., 2025).

However, we argue that predictive ability does not fully characterize the value of a representation. Although REP features contain adaptation-related information and may not be directly compatible with the fixed novel classifier, their optimization across multiple base categories may preserve category-level relationships and discriminative cues complementary to CLS features. Such information may not be explicitly retained in CLS features optimized for transferable recognition, but could still benefit novel-class recognition.

Therefore, instead of asking whether REP can directly predict novel categories, we study a different question: can REP provide useful information for CLS-based inference without serving as an independent predictor?

To answer this question, we first perform linear probe analysis and show that REP contains decodable information complementary to CLS features. Based on this observation, we develop a training-free and inference-efficient strategy, illustrated in Fig. 1(right), that exploits REP information while preserving the original inference pathway. When plugged into MMRL++, our method achieves state-of-the-art novel-class performance among methods without knowledge distillation.

2 Related Work

2.1 CLIP Adaptation and Adaptation-Specific Representations

CLIP adaptation methods differ substantially in how task-specific knowledge is introduced, but many share a common structural property: they contain learnable components that accumulate task information during adaptation. Prompt-learning methods instantiate such components as learnable textual or visual prompts (Zhou et al., 2022b;a; Khattak et al., 2023a), while other approaches introduce task-specific residuals, decomposed feature components, or dedicated representation tokens (Yu et al., 2023; Zhang et al., 2024; Guo & Gu, 2025; 2026). Recent methods further enrich adaptation with semantic or task-aware information from language models, class attributes, clustering, or other auxiliary sources (Zheng et al., 2024; Hao et al., 2025; Ding et al., 2025).

We use *REP* as a general term for representations produced by such learnable components that aggregate task-specific information. Accordingly, REP need not take a particular architectural form: it can be instantiated by dedicated representation tokens in MMRL/MMRL++, or by suitable visual prompt states in MaPLe and PromptSRC (Khattak et al., 2023a;b; Guo & Gu, 2025; 2026). This view allows REPURPOSE to operate across different adaptation frameworks by using their respective REP, without changing how the host model is originally trained.

2.2 Plug-in Enhancements for CLIP Adaptation

Beyond designing new adaptation architectures, another line of work improves existing CLIP adaptation methods through plug-in strategies. Some methods intervene during adaptation, for example by regulating optimization, introducing parallel prompting, or incorporating additional semantic and visual guidance (Li et al., 2023a; 2025; Chen et al., 2026; Yao et al., 2024; Cui et al., 2025; Zhou et al., 2026). Others leave the adapted model unchanged and refine its predictions at inference time.

DAC calibrates adapted predictions with knowledge from the pretrained model, while NeRP uses pretrained inter-class preferences together with sample-level evidence to resolve ambiguous predictions (Wang et al., 2024a; Tian et al., 2026).

In contrast, REPURPOSE operates on information already present in the adapted model.Rather than learning additional prompts or predictors, it reuses REP produced by the existing adaptation process, while frozen semantic prototypes are used to structure and verify the correction. This allows REPURPOSE to improve an existing host without retraining or altering its original adaptation procedure.

A broader comparison with adaptation, prediction-stage correction, and inference-time intervention methods is provided in Appendix G.

3 PRELIMINARIES AND MOTIVATION

3.1 PRELIMINARIES

We consider a CLIP-based vision–language model h adapted to a downstream task. Given an image x and label set $\mathcal{Y}$, let $z_{\mathrm{CLS}}(x) \in \mathbb{R}^d$ be the normalized final readout of the visual `[CLS]` token, and let $\{t_c^a\}_{c\in\mathcal{Y}}$ be the adapted text features. The host classifier computes

$$\ell_c(x) = \tau_{\mathrm{logit}} z_{\mathrm{CLS}}(x)^\top t_c^a, \qquad R_0(x) = \arg\max_{c\in\mathcal{Y}} \ell_c(x), \tag{1}$$

where τ_{logit} is the host logit scale. The host model and classifier remain fixed at inference.

The hosts considered here also contain visual tokens introduced during adaptation. We use *REP* to denote their normalized final readout,

$$z_{\mathrm{REP}}(x) \in \mathbb{R}^d. \tag{2}$$

Their final hidden states are mean-pooled, mapped through the CLS-aligned visual projection, and ℓ_2-normalized. Hence, z_{CLS} and z_{REP} come from different token pathways but lie in the same visual embedding space.

REP refers to the token states after processing the current image, not to their learnable input embeddings, and is therefore sample-dependent. The token construction differs across hosts; host-specific REP readouts and implementation details are provided in Appendix A.1.

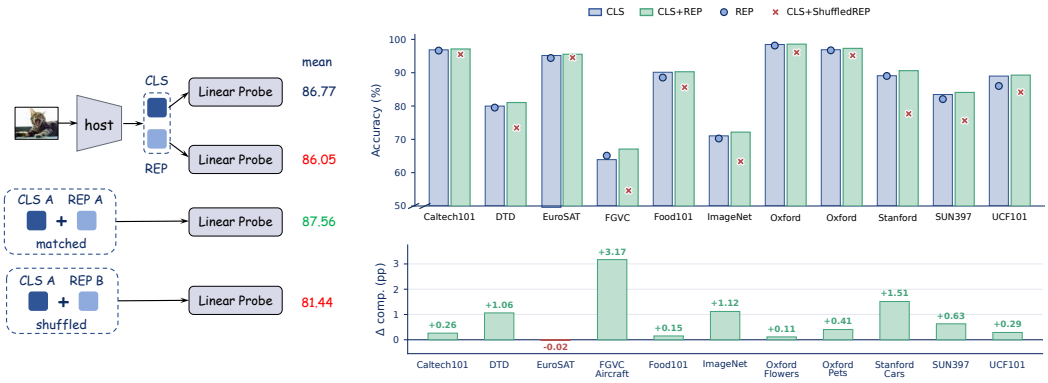

Figure 2: **Diagnostic linear-probe analysis of REP.** Left: linear probes on frozen CLS and REP representations. Right: novel-class probe performance across 11 datasets. REP remains linearly decodable and provides complementary information to CLS, while shuffling REP substantially degrades performance.

3.2 MOTIVATION

Does REP retain novel-class information? Although REP is not used by the host for novel-class prediction, it may still retain information that is poorly aligned with the fixed decision rule.

We test this using linear probing, a standard diagnostic for assessing whether class information is linearly decodable from a frozen representation (Alain & Bengio, 2017; Radford et al., 2021).

We freeze the adapted host and train identical linear classifiers on CLS, REP, and their matched concatenation. As a control, we shuffle REP across samples before concatenation, breaking its correspondence with the current image while preserving its distribution and dimensionality.

Figure 2 shows that REP reaches 86.05% mean probe accuracy, close to 86.77% for CLS. Matched CLS+REP improves this to 87.56% and outperforms CLS on 10 of the 11 datasets. Shuffling REP instead reduces the accuracy to 81.44%. REP therefore retains information that complements CLS and depends on the current sample.

Are host errors rank-local? Prior work has exploited small sets of top-ranked classes for refining CLIP predictions (Yi et al., 2024; Hakim et al., 2025). We therefore ask whether errors of base-to-novel adaptation hosts exhibit this locality systematically. For misclassified novel samples, we record the rank of the ground-truth class under MaPLe, PromptSRC, MMRL, and MMRL++.

The median ground-truth rank is 3 for all four hosts. Across all misclassified novel samples from the 11 datasets and three seeds, the ground-truth class remains within the host's top-5 predictions in 77.3–77.8% of cases.

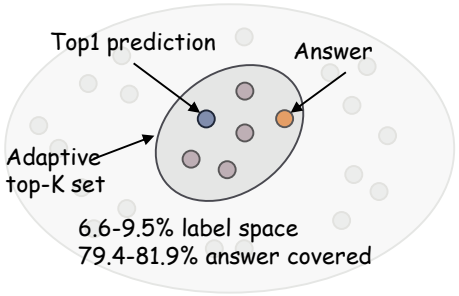

Figure 3: **Host errors are largely rank-local.** The entropy-adaptive candidate set retains only 6.6–9.5% of the label space while containing the ground-truth class for 79.4–81.9% of errors.

The entropy-adaptive candidate set used later covers 79.4–81.9% of error ground truths while retaining only 6.6–9.5% of the label space on average. On MMRL, it uses only 3.55 candidates on average while matching a fixed $K = 5$ in both error coverage (81.78% vs. 81.73%) and Novel accuracy (78.30% vs. 78.28%). Detailed cross-host statistics and fixed-K comparisons are provided in Appendices B.2 and B.4, respectively.

Host errors are therefore often *rank-local*: the top-1 prediction may be wrong while the correct class remains among a small set of high-ranked alternatives.

Does frozen CLIP retain corrective potential? Although adaptation improves downstream recognition, the original frozen CLIP may still correctly recognize samples that are misclassified by the adapted host. We examine this directly by revisiting the novel-class errors of MMRL and evaluating standard class-name zero-shot CLIP on exactly these errors. Across the 11 datasets and three seeds, frozen CLIP correctly recognizes 6,851 of the 43,100 novel-class errors made by MMRL (15.9%). This potential is also reflected in prior adaptation methods, which have repeatedly leveraged the original frozen CLIP as a source of auxiliary supervision or a training-time reference to preserve pretrained knowledge (Zhu et al., 2023; Khattak et al., 2023b; Guo & Gu, 2025; Yang et al., 2026).

4 METHOD

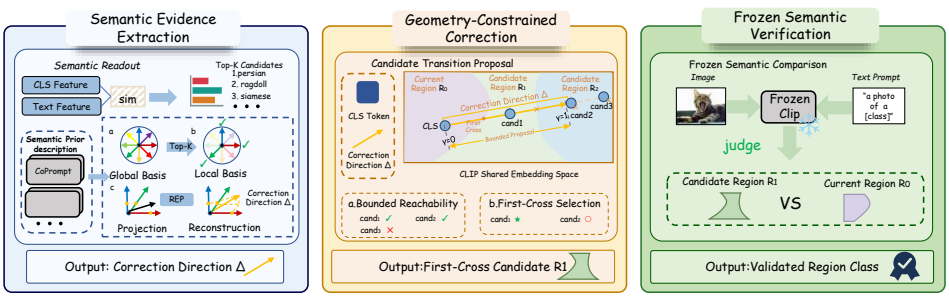

Figure 4: **Overview of REPURPOSE.** REP extracts a correction direction from semantic evidence; the host classifier determines the transition and frozen CLIP verifies it.

4.1 OVERVIEW

The three observations above directly motivate the design of REPURPOSE. First, since REP retains sample-specific information complementary to CLS, we use it as corrective evidence rather than as a standalone predictor. Second, since host errors are largely rank-local, we restrict correction to plausible alternatives around the original prediction. Third, since frozen CLIP retains corrective potential on host errors, we use its pretrained semantics to verify rather than generate corrections.

Figure 4 summarizes the resulting three-stage inference process. From left to right, **Semantic Evidence Extraction** reads the part of REP relevant to the current local ambiguity and produces a sample-specific correction direction $\delta(x)$. **Geometry-Constrained Correction** follows this direction from the original CLS representation under the unchanged host classifier, yielding a single candidate transition $c \to j^\star$. Finally, **Frozen Semantic Verification** checks whether pretrained CLIP semantics support this transition and either accepts it or retains the original prediction. Together, these stages implement a conservative correction principle: corrections are restricted to plausible local alternatives, favor the smallest REP-supported departure from the host decision, and are accepted only when independently supported by pretrained semantics.

4.2 SEMANTIC EVIDENCE EXTRACTION

The left panel of Fig. 4 answers a simple question: which part of REP is relevant to the current prediction ambiguity?

Candidate set. Since host errors are typically local, we restrict the readout to classes that remain plausible under the host prediction. Let $\pi = \mathrm{softmax}(\ell)$ be the host predictive distribution and $H(x)$ its entropy. We set the candidate-set size from the effective support of π,

$$K(x) = \min\Big\{C,\ \max\Big\{2,\ \Big\lceil e^{H(x)} \Big\rceil\Big\}\Big\}, \tag{3}$$

and denote the corresponding top-$K(x)$ classes by $\mathcal{C}_K(x)$. Confident samples therefore retain fewer alternatives, while uncertain samples retain more.

Semantic subspace. For each class, we encode a fixed set of textual descriptions with the frozen CLIP text encoder and average them into a normalized semantic prototype p_c. These prototypes are used only to describe class-level semantic structure; they do not replace the adapted text features t_c^a. Details of the semantic-prototype construction are provided in Appendix A.3.

We first extract the dominant semantic variation across all class prototypes to form a global semantic basis. The prototypes of the current candidates then define a local subspace within this basis. Let $B(x)$ denote an orthonormal basis of this candidate-induced semantic subspace. We use the same 95% semantic-energy criterion for the global basis across all datasets and hosts; performance remains stable across a range of energy thresholds (Appendix B.6). Detailed constructions of the global and instance-adaptive local bases are provided in Appendices A.4 and A.5.

REP readout. We retain only the component of REP that lies in the current semantic subspace:

$$\delta(x) = B(x)B(x)^\top z_{\mathrm{REP}}. \tag{4}$$

Thus, $\delta(x)$ keeps the part of REP aligned with the semantic variation among the current candidates and removes components unrelated to this local ambiguity. It is not converted into class logits and does not produce a prediction. Its only role is to define a correction direction.

4.3 GEOMETRY-CONSTRAINED CORRECTION

The semantic readout provides a direction, but not a target class. We therefore leave the target entirely to the geometry of the unchanged host classifier.

Starting from the original CLS representation, we consider the bounded trajectory

$$z(\gamma) = z_{\mathrm{CLS}} + \gamma\,\delta(x), \qquad 0 < \gamma \le 1. \tag{5}$$

The bound prevents the correction from extrapolating beyond the REP evidence that has been extracted.

Bounded reachability. Let $c = R_0(x)$ be the original prediction. For a candidate $j \in \mathcal{C}_K(x)\backslash\{c\}$, the step size required to reach its pairwise decision boundary along $\delta(x)$ is

$$\gamma_j = \frac{z_{\mathrm{CLS}}^\top (t_c^a - t_j^a)}{\delta(x)^\top (t_j^a - t_c^a)}. \tag{6}$$

A candidate is reachable when the correction direction moves toward that candidate and its boundary is crossed within $0 < \gamma_j \leq 1$. We denote the set of such candidates by $\mathcal{R}(x)$.

First-cross selection. Among the reachable candidates, we take the boundary encountered first:

$$j^\star = \arg\min_{j \in \mathcal{R}(x)} \gamma_j. \tag{7}$$

If $\mathcal{R}(x)$ is empty, the original prediction is kept.

The first-cross rule follows a minimum-intervention principle: among candidate transitions reachable along the REP-induced direction, it selects the one requiring the smallest departure from the original host representation. A geometric derivation and the directional, candidate-restricted minimum-intervention interpretation of the first-cross rule are provided in Appendix F.

4.4 Frozen Semantic Verification

Geometry produces a single proposal $c \to j^\star$. Since the correction direction comes from an adapted representation, we do not accept this transition automatically. Instead, frozen CLIP provides a final semantic check.

Let

$$v(x) = \mathrm{norm}(g_v^f(x))$$

be the frozen CLIP image representation. We measure the relative semantic support for the proposed transition by

$$F_{c \to j^\star}(x) = v(x)^\top \big(p_{j^\star} - p_c\big). \tag{8}$$

A positive value means that frozen CLIP assigns higher image–text similarity to $j^\star$ than to the original class c. In this case we accept the correction; otherwise, we retain the original host prediction.

Frozen CLIP therefore serves only as a verifier of the host-proposed transition, rather than as an alternative classifier. To avoid the additional frozen-CLIP image forward, we also consider a lightweight adapted verifier that replaces $v(x)$ with the already available host representation $z_{\mathrm{CLS}}(x)$:

$$F^a_{c \to j^\star}(x) = z_{\mathrm{CLS}}(x)^\top \big(p_{j^\star} - p_c\big). \tag{9}$$

The proposal $j^\star$, frozen semantic prototypes, and acceptance rule remain unchanged.

5 Experiments

5.1 Experimental Setup

Protocol and hosts. We follow the standard 16-shot base-to-novel protocol (Zhou et al., 2022b; Khattak et al., 2023a) on 11 benchmarks: ImageNet, Caltech101, OxfordPets, StanfordCars, Flowers102, Food101, FGVCAircraft, SUN397, DTD, EuroSAT, and UCF101 (Deng et al., 2009; Fei-Fei et al., 2004; Parkhi et al., 2012; Krause et al., 2013; Nilsback & Zisserman, 2008; Bossard et al., 2014; Maji et al., 2013; Xiao et al., 2010; Cimpoi et al., 2014; Helber et al., 2019; Soomro et al., 2012). We evaluate REPurpose on four CLIP-based hosts using their officially released checkpoints: MaPLe (Khattak et al., 2023a), PromptSRC (Khattak et al., 2023b), MMRL (Guo & Gu, 2025), and MMRL++ (Guo & Gu, 2026). REPurpose leaves the trained host and its classifier unchanged.

Results are averaged over three seeds. For base-to-novel evaluation, Base and Novel accuracies are macro-averaged over the 11 datasets, and HM is computed from these two averages. Controlled analyses use the official MMRL checkpoints, visual REP, and the publicly released class-level textual descriptions from CoPrompt for constructing semantic prototypes (Roy & Etemad, 2024). Host-specific readouts and implementation details are in Appendix A.1.

Table 1: **Base-to-novel generalization across 11 datasets.** "•" indicates that the original host prediction is retained on base classes. Green (red) arrows show increases (decreases) in Novel accuracy relative to the corresponding host. Bold and underline denote the best and second-best numeric results, respectively.

Method	Average			ImageNet			Caltech101			OxfordPets			StanfordCars			Flowers102		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP (ICML 21)	69.34	74.22	71.70	72.43	68.14	70.22	96.84	94.00	95.40	91.17	97.26	94.12	63.37	74.89	68.65	72.08	77.80	74.83
CoOp (IJCV 22)	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47	78.12	60.40	68.13	97.60	59.67	74.06
CoCoOp (CVPR 22)	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43	70.49	73.59	72.01	94.87	71.75	81.71
LLaMP (CVPR 24)	85.16	77.71	81.27	77.99	71.27	74.48	98.45	95.85	97.13	96.31	97.74	97.02	81.56	74.54	77.89	97.82	77.40	86.42
CoPrompt (ICLR 24)	84.00	77.23	80.48	77.67	71.27	74.33	98.27	94.90	96.55	95.67	98.10	96.87	76.97	74.40	75.66	97.27	76.60	85.71
TAC (CVPR 25)	85.24	77.60	81.24	78.57	71.03	74.61	98.57	95.27	96.89	95.93	**98.17**	97.04	81.63	74.17	77.72	97.97	76.87	86.15
TAP (ICLR 25)	84.75	77.63	81.04	77.97	70.40	73.99	98.90	95.50	97.17	95.80	97.73	96.76	80.70	74.27	77.35	97.90	75.57	85.30
ManiPT (ICML 26)	**85.82**	78.67	82.09	**78.67**	72.01	**75.20**	98.95	**95.94**	**97.42**	**96.52**	98.15	**97.33**	**82.53**	75.61	**78.92**	98.12	76.84	86.19
NeRP (ICML 26)	85.68	78.04	81.68	77.90	71.53	74.58	98.97	95.10	97.00	95.90	97.70	96.79	81.30	75.30	78.19	**98.97**	**78.43**	**87.51**
MaPLe (CVPR 23)	82.28	75.14	78.55	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58	72.94	74.00	73.47	95.92	72.46	82.56
+REPurpose (Ours)	•	76.68 ↑1.54	79.38	•	71.64 ↑1.10	74.07	•	95.85 ↑1.49	96.79	•	97.58 ↓0.18	96.49	•	75.00 ↑1.00	73.96	•	74.89 ↑2.43	84.11
PromptSRC (ICCV 23)	84.26	76.10	79.97	77.60	70.73	74.01	98.10	94.03	96.02	95.33	97.30	96.30	78.27	74.97	76.58	98.07	76.50	85.95
+REPurpose (Ours)	•	77.41 ↑1.31	80.69	•	71.75 ↑1.02	74.56	•	95.52 ↑1.49	96.79	•	97.35 ↑0.05	96.33	•	75.89 ↑0.92	77.06	•	77.12 ↑0.62	86.34
MMRL (CVPR 25)	85.68	77.17	81.20	77.90	71.43	74.53	98.97	94.47	96.67	95.90	97.58	96.73	81.30	75.09	78.07	**98.97**	77.26	86.78
+REPurpose (Ours)	•	78.30 ↑1.13	81.82	•	72.55 ↑1.12	75.13	•	95.71 ↑1.24	97.31	•	97.82 ↑0.24	96.85	•	75.92 ↑0.83	78.52	•	77.38 ↑0.12	86.85
MMRL++ (IJCV 26)	85.53	78.32	81.77	77.63	71.50	74.44	**99.07**	94.54	96.75	95.60	97.45	96.51	81.33	75.24	78.17	98.53	77.92	87.02
+REPurpose (Ours)	•	**79.25** ↑0.93	**82.27**	•	**72.73** ↑1.23	75.10	•	95.60 ↑1.06	97.30	•	97.95 ↑0.50	96.76	•	**76.06** ↑0.82	78.61	•	77.61 ↓0.31	86.83

Method	Food101			FGVCAircraft			SUN397			DTD			EuroSAT			UCF101		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP (ICML 21)	90.10	91.22	90.66	27.19	36.29	31.09	69.36	75.35	72.23	53.24	59.90	56.37	56.48	64.05	60.03	70.53	77.50	73.85
CoOp (IJCV 22)	88.33	82.26	85.19	40.44	22.30	28.75	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
CoCoOp (CVPR 22)	90.70	91.29	90.99	33.41	23.71	27.74	79.74	76.86	78.27	77.01	56.00	64.85	87.49	60.04	71.21	82.33	73.45	77.64
LLaMP (CVPR 24)	91.05	91.93	91.49	47.30	37.61	41.90	83.41	79.90	81.62	83.49	64.49	72.77	91.93	83.66	87.60	87.13	80.66	83.77
CoPrompt (ICLR 24)	90.73	92.07	91.40	40.20	39.33	39.76	82.63	80.03	81.31	83.13	64.73	72.79	94.60	78.57	85.84	86.90	79.57	83.07
TAC (CVPR 25)	90.87	91.87	91.37	44.60	37.70	40.86	**83.70**	80.03	**81.82**	83.37	64.27	72.58	94.37	82.60	88.10	88.07	81.67	84.75
TAP (ICLR 25)	90.97	91.83	91.40	44.40	36.50	40.06	82.87	79.53	81.17	84.20	68.00	75.24	90.70	82.17	86.22	87.90	**82.43**	85.08
ManiPT (ICML 26)	**91.17**	**92.16**	**91.66**	**48.28**	39.87	**43.68**	83.18	80.18	81.65	84.38	68.56	75.65	94.06	83.72	88.59	**88.14**	82.31	**85.13**
NeRP (ICML 26)	90.57	91.55	91.06	46.30	37.67	41.54	83.20	79.60	81.36	**85.67**	65.70	74.37	95.60	85.17	90.08	88.10	80.67	84.22
MaPLe (CVPR 23)	90.71	92.05	91.38	37.44	35.61	36.50	80.82	78.70	79.75	80.36	59.18	68.16	94.07	73.23	82.35	83.00	78.66	80.77
+REPurpose (Ours)	•	92.12 ↑0.07	91.41	•	37.41 ↑1.80	37.43	•	79.87 ↑1.17	80.34	•	65.02 ↑5.84	71.88	•	74.29 ↑1.06	83.01	•	79.79 ↑1.13	81.36
PromptSRC (ICCV 23)	90.67	91.53	91.10	42.73	37.87	40.15	82.67	78.47	80.52	83.37	62.97	71.75	92.90	73.90	82.32	87.10	78.80	82.74
+REPurpose (Ours)	•	91.81 ↑0.28	91.24	•	39.53 ↑1.66	41.07	•	79.39 ↑0.92	81.00	•	67.87 ↑4.90	74.83	•	74.82 ↑0.92	82.89	•	80.46 ↑1.66	83.65
MMRL (CVPR 25)	90.57	91.52	91.04	46.30	37.05	41.16	83.20	79.29	81.20	**85.67**	64.94	73.88	95.60	80.15	87.20	88.10	80.06	83.89
+REPurpose (Ours)	•	91.73 ↑0.21	91.15	•	39.39 ↑2.34	42.57	•	80.07 ↑0.78	81.60	•	68.32 ↑3.38	76.02	•	80.61 ↑0.46	87.47	•	81.77 ↑1.71	84.82
MMRL++ (IJCV 26)	90.47	91.74	91.10	46.40	38.75	42.23	83.03	79.59	81.27	85.47	66.02	74.50	**95.93**	88.22	91.91	87.37	80.58	83.84
+REPurpose (Ours)	•	91.89 ↑0.15	91.17	•	**40.05** ↑1.30	42.99	•	**80.33** ↑0.74	81.66	•	**69.20** ↑3.18	**76.48**	•	**88.23** ↑0.01	**91.92**	•	82.13 ↑1.55	84.67

Transfer and comparison protocols. For cross-dataset (CD) evaluation, ImageNet-adapted MMRL and MMRL++ models are evaluated directly on the other ten datasets. For domain generalization (DG), we evaluate ImageNet-V2, ImageNet-Sketch, ImageNet-A, and ImageNet-R (Recht et al., 2019; Wang et al., 2019; Hendrycks et al., 2021b;a) without target-domain training or calibration. We report unweighted means over the ten CD targets and four DG domains. For NeRP (Tian et al., 2026), we use the published results for base-to-novel evaluation and our reproduction for subsequent experiments, using the authors' released code and inference configurations without target-test tuning. Table 1 additionally includes LLaMP (Zheng et al., 2024), TAC (Hao et al., 2025), TAP (Ding et al., 2025), and ManiPT (Yang et al., 2026). All deltas are computed from values rounded to two decimal places.

5.2 Base-to-Novel Generalization

Table 1 shows that REPurpose consistently improves novel-class generalization while leaving Base predictions unchanged. When applied to MMRL++, REPurpose achieves an average Novel accuracy of **79.25%**, demonstrating performance competitive with recent state-of-the-art methods. The improvement is particularly pronounced on challenging benchmarks such as DTD and FGVCAircraft, where REPurpose yields consistent gains across different hosts, with improvements of up to **+5.84** and **+2.34** points, respectively. Importantly, the gains persist even on the strong MMRL++ baseline, improving its average Novel accuracy from 78.32% to **79.25%** while preserving its Base predictions. Paired three-seed statistics are reported in Appendix B.8.

When can correction hurt? EuroSAT is particularly challenging for frozen CLIP, as also observed in prior work on satellite imagery (Radford et al., 2021; Jain et al., 2025). This makes frozen CLIP a less reliable verifier for the strongly adapted MMRL++ host, for which the original correction reduces Novel accuracy from 88.22% to 86.31%. We therefore introduce a simple label-free safeguard that increases the verification threshold when the adapted host and frozen CLIP show a large novel-response mismatch. This recovers the accuracy to 88.23%. We use a fixed scale of 0.07,

Table 2: **Controlled analysis on MMRL.** (a) Component ablation of REPURPOSE. (b) REP information controls. Direct fusion tests whether REP provides complementary evidence to CLS, while shuffling breaks the image–REP correspondence with the remaining inference pipeline unchanged.

(a) Component ablation

REP Corr.	Global	Local	Verify	Novel Acc.
				77.17
✓			✓	77.86
✓	✓		✓	77.95
✓	✓	✓		77.29
✓	✓	✓	✓	**78.30**

(b) REP information controls

Use of REP	Novel Acc.	Δ
CLS only	77.17	–
Raw REP fusion	77.26	+0.09
Projected REP fusion	77.70	+0.53
Shuffled REPURPOSE	77.40	+0.23
Matched REPURPOSE	**78.30**	**+1.13**

following CLIP's initial temperature, as both serve to set the scale of margins in the normalized image–text similarity space; Appendix C shows that the result is insensitive to this choice.

5.3 CONTROLLED ANALYSIS

Component ablation. Table 2(a) starts from the 77.17% CLS baseline. Using raw REP as the correction direction with verification reaches 77.86%. Global projection gives 77.95%, and local projection further improves accuracy to 78.30%. Removing verification from the full model reduces accuracy to 77.29%.

REP information controls. Table 2(b) compares REPURPOSE with simpler uses of REP. Direct fusion with raw and projected REP reaches 77.26% and 77.70%, respectively. Shuffling REP across samples reduces REPURPOSE from 78.30% to 77.40%, showing that the correction relies on the REP of the current image. Together, these controls support using REP as sample-specific corrective evidence rather than as a feature directly fused with CLS. For fusion, we report the best $\lambda \in \{0.25, 0.5, 1.0\}$; the full sweep is in Appendix B.3.

Why the first crossing? We compare successive reachable crossings of the REP-induced trajectory, aggregating MMRL and MMRL++ over 11 datasets and three seeds under the same verifier. The first crossing achieves the highest macro accuracy (78.69%) and correction precision (51.89%). At the second and third crossings, precision drops to 24.87% and 14.42%, while the ground-truth hit rate falls from 14.98% to 2.33% and 0.60%. These results suggest that useful corrective information is concentrated at the earliest decision transition, supporting the first-cross rule in REPURPOSE. Full statistics are provided in Appendix B.7.

Does the gain come from richer semantics? CoPrompt descriptions clearly contribute to performance, but they do not account for the gains from REP. Under matched CoPrompt semantics, adding REP improves MMRL from 77.78% to 78.30% and MMRL++ from 77.97% to 79.25%. The effect is even larger with standard class-name semantics (75.15%→77.37% and 75.66%→78.30%). Thus, richer semantic descriptions improve the final system, while REP provides complementary corrective evidence under either semantic resource. See Appendix B.5 for further analysis.

Table 3: **REP vs. semantic resource.** Novel accuracy (%) averaged over 11 datasets.

Host	Evidence	Name	CoPrompt
MMRL	Semantic-only	75.15	77.78
	+ REP	77.37	**78.30**
MMRL++	Semantic-only	75.66	77.97
	+ REP	78.30	**79.25**

5.4 COMPARISON AND TRANSFER

Table 4 reports cross-dataset transfer. Applying REPURPOSE only on target datasets raises MMRL from 67.26% to 68.14% and MMRL++ from 67.31% to 68.22%. On domain generalization, the corresponding averages increase from 60.52% to 61.19% and from 60.42% to 61.06%; both hosts improve on all four shifts. Per-domain results are in Appendix D.1.

Efficiency and configuration portability. As shown in Figure 5 (left), under the matched MMRL setting, both verifier variants of REPURPOSE achieve higher accuracy than NeRP across B2N, CD,

Table 4: **Cross-dataset transfer from an ImageNet source model.** "•" indicates that the source prediction is unchanged; REPURPOSE is applied only on target datasets. Bold denotes the best result in each column.

Method	Source	Target										
	Image Net	Caltech 101	Oxford Pets	Stanford Cars	Flowers 102	Food 101	FGVC Aircraft	SUN 397	DTD	EuroSAT	UCF101	Avg.
CoOp (IJCV 22)	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
CoCoOp (CVPR 22)	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
MaPLe (CVPR 23)	70.72	**95.53**	90.49	65.57	72.20	86.20	24.74	67.01	46.49	48.06	68.69	66.50
CoPrompt (ICLR 24)	70.80	94.50	90.73	65.67	72.30	86.43	24.00	67.57	47.07	51.90	69.73	66.99
TAC (CVPR 25)	72.77	94.53	90.67	65.30	72.20	85.83	23.53	67.63	47.57	48.07	70.00	66.53
TAP (ICLR 25)	72.30	94.30	90.70	65.60	70.93	86.10	24.57	68.30	50.20	46.00	68.90	66.56
ManiPT (ICML 26)	**72.92**	95.13	90.87	66.27	72.14	86.66	27.00	68.50	**51.83**	50.88	**71.11**	68.04
MMRL (CVPR 25)	72.00	94.56	91.37	65.78	72.47	86.35	26.38	67.55	46.36	53.01	68.75	67.26
+REPURPOSE (Ours)	•	94.50	**92.56**	66.78	72.45	86.57	**28.06**	68.79	48.50	53.57	69.61	68.14
MMRL++ (IJCV 26)	71.77	94.51	91.12	66.19	**73.14**	86.52	25.78	67.67	45.63	53.33	69.20	67.31
+REPURPOSE (Ours)	•	94.69	92.39	**67.05**	72.43	**86.68**	26.77	**68.98**	48.96	**54.30**	70.00	**68.22**

Verifier comparison

Host	Setting	B2N	CD	DG
MMRL	REPURPOSE (Adapted)	**78.33**	68.09	61.11
	REPURPOSE (Frozen)	78.30	**68.14**	**61.19**
	NeRP	77.38	67.77	60.83
MMRL++	REPURPOSE (Adapted)	79.13	**68.24**	61.04
	REPURPOSE (Frozen)	**79.25**	68.22	**61.06**

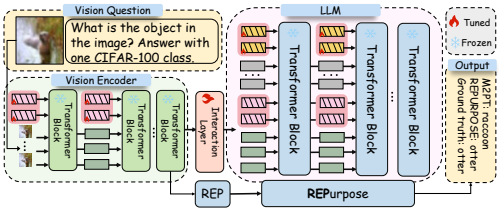

Figure 5: **REPURPOSE verifier comparison and extension to autoregressive inference.** Left: adapted and frozen verifiers across B2N, CD, and DG, with NeRP shown under the matched MMRL setting. Right: integration of REPURPOSE into the M2PT autoregressive inference pipeline.

and DG. Moreover, while NeRP relies on dataset-specific inference configurations, REPURPOSE uses a single set of method-specific hyperparameters across datasets without per-dataset tuning. In addition, REPURPOSE is more efficient: NeRP requires $1.69\times$ runtime (3.013 ms/image), whereas REPURPOSE requires only $1.17\times$ (2.020 ms/image) with the adapted verifier and $1.35\times$ (2.333 ms/image) with the frozen verifier. The two verifiers perform similarly across B2N, CD, and DG, differing by at most 0.12 pp, making the adapted verifier a lightweight option for practical deployment. Full settings are in Appendix D.3.

Autoregressive extension. As illustrated in Figure 5 (right), REPURPOSE extends naturally to autoregressive inference. On CIFAR-100, it improves M2PT (Wang et al., 2024b) from 59.21% to 60.62% without modifying autoregressive decoding. Details are provided in Appendix E.

6 CONCLUSION

We introduce REPURPOSE, a training-free inference framework that repurposes adaptation-specific representations as corrective evidence for novel-class prediction. Without modifying host training or parameters, REPURPOSE improves novel-class generalization across multiple adaptation methods and enables MMRL++ to achieve state-of-the-art novel performance among methods without knowledge distillation. The same correction principle also extends to the autoregressive M2PT (Wang et al., 2024b) pipeline without modifying decoding.

Our results also suggest a natural direction for future adaptation. MMRL already incorporates the base-class role of REP into training by applying a dedicated classification objective to its representation features, and then uses REP jointly with CLS at base inference (Guo & Gu, 2025). In contrast, REPURPOSE reveals a new role for REP at novel inference while leaving its training unchanged. A natural next step is therefore to incorporate this corrective role into the training objective, explicitly optimizing REP to provide better correction directions for novel inference.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting and editing, LaTeX and table formatting, and discussions of experimental organization, ablation design, and interpretation of empirical results. All experimental runs and numerical results were produced independently of the generative AI system, and all AI-assisted text, mathematical statements, implementation descriptions, and scientific claims were reviewed and verified by the authors before inclusion. The authors take full responsibility for the content of the final manuscript.

ETHICS STATEMENT

This work studies inference-time correction for vision-language models using standard publicly available image-classification benchmarks and pretrained models. It does not involve human subjects, private or personally identifiable information, or the release of new sensitive datasets. As with other image-classification systems built on large pretrained models, downstream use may inherit biases, representation gaps, or failure modes from the underlying models and training data. We do not claim that REPURPOSE removes these risks, and deployment in consequential settings would require task-specific validation beyond the benchmarks studied here.

REPRODUCIBILITY STATEMENT

We describe the complete inference procedure and all method-specific definitions in Sec. 4, including construction of the global and local semantic spaces, entropy-adaptive candidate selection, the bounded first-crossing rule, and frozen-CLIP verification. The semantic prototypes used by REPURPOSE are constructed from the released CoPrompt descriptions (Roy & Etemad, 2024), as detailed in Appendix A.3. We use the same method settings across datasets and hosts and do not tune method-specific hyperparameters for individual datasets.

REPURPOSE is training-free. For the main experiments, we directly apply it to the released checkpoints of MaPLe (Khattak et al., 2023a), PromptSRC (Khattak et al., 2023b), MMRL (Guo & Gu, 2025), and MMRL++ (Guo & Gu, 2026), without retraining, fine-tuning, or modifying their parameters. Host-specific REP readouts are detailed in Appendix A.1. Experimental datasets, checkpoint usage, random seeds, transfer protocols, aggregation conventions, and controlled-analysis settings are specified in Sec. 5.1. Controlled and transfer experiments use paired evaluations of the same host checkpoints whenever a correction method is compared with its baseline. Additional transfer results and seed-level statistics are reported in Appendices D and B.8, respectively. The geometric derivation of the first-cross correction is provided in Appendix F.

An anonymized implementation will be provided with the submission as supplementary material, together with configuration files, evaluation scripts, and instructions for reproducing the main base-to-novel, controlled-analysis, cross-dataset, and domain-generalization experiments. The release will also include scripts for the semantic-resource, fusion, candidate-set, semantic-energy, crossing-order, verifier, seed-stability, and runtime analyses. For our NeRP reproduction, we will additionally provide the released configuration settings, evaluation scripts, paired-checkpoint protocol, and seed-level outputs. Reproduced external results are explicitly distinguished from numbers reported by the original papers.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, volume 8694 of *Lecture Notes in Computer Science*, pp. 446–461. Springer, 2014. doi: 10.1007/978-3-319-10599-4_29. URL https://doi.org/10.1007/978-3-319-10599-4_29.

Xueting Chen, Jun-Jie Huang, Yan Yan, Long Lan, Yuhua Tang, and Wenjing Yang. Do-prompt: Causal interventions meet variational prompt bottlenecks. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=QW3mKUstBr.

Arpita Chowdhury, Dipanjyoti Paul, Zheda Mai, Jianyang Gu, Ziheng Zhang, Kazi Sajeed Mehrab, Elizabeth G. Campolongo, Daniel I. Rubenstein, Charles V. Stewart, Anuj Karpatne, Tanya Y. Berger-Wolf, Yu Su, and Wei-Lun Chao. Prompt-cam: Making vision transformers interpretable for fine-grained analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 4375–4385. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00413. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chowdhury_Prompt-CAM_Making_Vision_Transformers_Interpretable_for_Fine-Grained_Analysis_CVPR_2025_paper.html.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Th6NyL07na.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 3606–3613. IEEE Computer Society, 2014. doi: 10.1109/CVPR.2014.461. URL https://doi.org/10.1109/CVPR.2014.461.

Fangming Cui, Yonggang Zhang, Xuan Wang, Xinmei Tian, and Jun Yu. Enhancing target-unspecific tasks through a features matrix. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/cui25g.html.

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 Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Tong Ding, Wanhua Li, Zhongqi Miao, and Hanspeter Pfister. Tree of attributes prompt learning for vision-language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=wFs2E5wCw6.

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 *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2004, Washington, DC, USA, June 27 - July 2, 2004*, pp. 178. IEEE Computer Society, 2004. doi: 10.1109/CVPR.2004.383. URL https://doi.org/10.1109/CVPR.2004.383.

Yuncheng Guo and Xiaodong Gu. MMRL: multi-modal representation learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 25015–25025. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02329. URL https://openaccess.thecvf.com/content/CVPR2025/html/Guo_MMRL_Multi-Modal_Representation_Learning_for_Vision-Language_Models_CVPR_2025_paper.html.

Yuncheng Guo and Xiaodong Gu. MMRL++: parameter-efficient and interaction-aware representation learning for vision-language models. *Int. J. Comput. Vis.*, 134(1):11, 2026. doi: 10.1007/S11263-025-02586-1. URL https://doi.org/10.1007/s11263-025-02586-1.

Gustavo Adolfo Vargas Hakim, David Osowiechi, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Moslem Yazdanpanah, Ismail Ben Ayed, and Christian Desrosiers. Clipartt: Adaptation of CLIP to new domains at test time. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2025, Tucson, AZ, USA, February 26 - March 6, 2025*, pp. 7092–7101. IEEE, 2025. doi: 10.1109/WACV61041.2025.00689. URL https://doi.org/10.1109/WACV61041.2025.00689.

Fusheng Hao, Fengxiang He, Fuxiang Wu, Tichao Wang, Chengqun Song, and Jun Cheng. Task-aware clustering for prompting vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 14745–14755. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01374. URL https://openaccess.thecvf.com/content/CVPR2025/html/Hao_Task-Aware_Clustering_for_Prompting_Vision-Language_Models_CVPR_2025_paper.html.

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 J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 12(7):2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242. URL https://doi.org/10.1109/JSTARS.2019.2918242.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 8320–8329. IEEE, 2021a. doi: 10.1109/ICCV48922.2021.00823. URL https://doi.org/10.1109/ICCV48922.2021.00823.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 15262–15271. Computer Vision Foundation / IEEE, 2021b. doi: 10.1109/CVPR46437.2021.01501. URL https://openaccess.thecvf.com/content/CVPR2021/html/Hendrycks_Natural_Adversarial_Examples_CVPR_2021_paper.html.

Pallavi Jain, Dino Ienco, Roberto Interdonato, Tristan Berchoux, and Diego Marcos. Senclip: Enhancing zero-shot land-use mapping for sentinel-2 with ground-level prompting. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2025, Tucson, AZ, USA, February 26 - March 6, 2025*, pp. 5656–5665. IEEE, 2025. doi: 10.1109/WACV61041.2025.00552. URL https://doi.org/10.1109/WACV61041.2025.00552.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 25004–25014. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02328. URL https://openaccess.thecvf.com/content/CVPR2025/html/Jiang_Devils_in_Middle_Layers_of_Large_Vision-Language_Models_Interpreting_Detecting_CVPR_2025_paper.html.

Muhammad Uzair Khattak, Hanoona Abdul Rasheed, Muhammad Maaz, Salman H. Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pp. 19113–19122. IEEE, 2023a. doi: 10.1109/CVPR52729.2023.01832. URL https://doi.org/10.1109/CVPR52729.2023.01832.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15190–15200, October 2023b.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pp. 554–561. IEEE Computer Society, 2013. doi: 10.1109/ICCVW.2013.77. URL https://doi.org/10.1109/ICCVW.2013.77.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13872–13882. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01316. URL https://doi.org/10.1109/CVPR52733.2024.01316.

Haoyang Li, Liang Wang, Chao Wang, Jing Jiang, Yan Peng, and Guodong Long. DPC: dual-prompt collaboration for tuning vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 25623–25632. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02386. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_DPC_Dual-Prompt_Collaboration_for_Tuning_Vision-Language_Models_CVPR_2025_paper.html.

Juncheng Li, Minghe Gao, Longhui Wei, Siliang Tang, Wenqiao Zhang, Mengze Li, Wei Ji, Qi Tian, Tat-Seng Chua, and Yueting Zhuang. Gradient-regulated meta-prompt learning for generalizable vision-language models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 2551–2562. IEEE, 2023a. doi: 10.1109/ICCV51070.2023.00241. URL https://doi.org/10.1109/ICCV51070.2023.00241.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023b. URL http://papers.nips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Qi Ma, Chen-Yang Wang, Dehong Gao, and Deng-Ping Fan. Alignednorm: Prompting vision–language models via coupled prompt field. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=aQAWAtrxxe.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. URL http://arxiv.org/abs/1306.5151.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008*, pp. 722–729. IEEE Computer Society, 2008. doi: 10.1109/ICVGIP.2008.47. URL https://doi.org/10.1109/ICVGIP.2008.47.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pp. 3498–3505. IEEE Computer Society, 2012. doi: 10.1109/CVPR.2012.6248092. URL https://doi.org/10.1109/CVPR.2012.6248092.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 2019. URL http://proceedings.mlr.press/v97/recht19a.html.

Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=wsRXwlwx4w.

Anushka Sivakumar, Andrew Zhang, Zaber Ibn Abdul Hakim, and Christopher Thomas. Steervlm: Robust model control through lightweight activation steering for vision language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 23640–23665. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-EMNLP.1285. URL https://doi.org/10.18653/v1/2025.findings-emnlp.1285.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. URL http://arxiv.org/abs/1212.0402.

Senmao Tian, Xiang Wei, and Shunli Zhang. Neutral-reference prompting for vision–language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=F5XJeuhwU7.

Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 10506–10518, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3eefceb8087e964f89c2d59e8a249915-Abstract.html.

Shuoyuan Wang, Jindong Wang, Guoqing Wang, Bob Zhang, Kaiyang Zhou, and Hongxin Wei. Open-vocabulary calibration for fine-tuned CLIP. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51734–51754. PMLR / OpenReview.net, 2024a. URL https://proceedings.mlr.press/v235/wang24bw.html.

Taowen Wang, Yiyang Liu, James Liang, Junhan Zhao, Yiming Cui, Yuning Mao, Shaoliang Nie, Jiahao Liu, Fuli Feng, Zenglin Xu, Cheng Han, Lifu Huang, Qifan Wang, and Dongfang Liu. M^2pt: Multimodal prompt tuning for zero-shot instruction learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 3723–3740. Association for Computational Linguistics, 2024b. doi: 10.18653/V1/2024.EMNLP-MAIN.218. URL https://doi.org/10.18653/v1/2024.emnlp-main.218.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June*

2010, pp. 3485–3492. IEEE Computer Society, 2010. doi: 10.1109/CVPR.2010.5539970. URL https://doi.org/10.1109/CVPR.2010.5539970.

Xi Yang, Yuanrong Xu, Weigang Zhang, Guangming Lu, David Zhang, and Jie Wen. Prompt tuning for CLIP on the pretrained manifold. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=ROFqW86Dca.

Hantao Yao, Rui Zhang, and Changsheng Xu. TCP: textual-based class-aware prompt tuning for visual-language model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 23438–23448. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02212. URL https://doi.org/10.1109/CVPR52733.2024.02212.

Chao Yi, Lu Ren, De-Chuan Zhan, and Han-Jia Ye. Leveraging cross-modal neighbor representation for improved CLIP classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 27392–27401. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02587. URL https://doi.org/10.1109/CVPR52733.2024.02587.

Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 10899–10909. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01049. URL https://doi.org/10.1109/CVPR52729.2023.01049.

Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 12924–12933. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01228. URL https://doi.org/10.1109/CVPR52733.2024.01228.

Zhaoheng Zheng, Jingmin Wei, Xuefeng Hu, Haidong Zhu, and Ram Nevatia. Large language models are good prompt learners for low-shot image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 28453–28462. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02688. URL https://doi.org/10.1109/CVPR52733.2024.02688.

Jingtao Zhou, Xirui Kang, feiyang huang, and Lai Man Po. SpecPL: Disentangling spectral granularity for prompt learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=118JD1BQxO.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 16795–16804. IEEE, 2022a. doi: 10.1109/CVPR52688.2022.01631. URL https://doi.org/10.1109/CVPR52688.2022.01631.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022b. doi: 10.1007/S11263-022-01653-1. URL https://doi.org/10.1007/s11263-022-01653-1.

Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 15613–15623. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01435. URL https://doi.org/10.1109/ICCV51070.2023.01435.

CONTENTS

A Method and Implementation Details **17**

A.1 Host-Specific REP Readout 17

A.2 Implementation Configuration 17

A.3 Semantic Prototype Construction 18

A.4 Global Semantic Basis 18

A.5 Instance-Adaptive Local Semantic Basis 19

A.6 Inference Procedure 20

B Additional Analyses **21**

B.1 Additional Motivation Diagnostics 21

B.2 Detailed Analysis of Rank-Local Host Errors 22

B.3 Alternative Inference Controls 23

B.4 Candidate-Set Size 23

B.5 Semantic Resource Ablation 24

B.6 Semantic Energy Sensitivity 25

B.7 Crossing-Order Analysis 25

B.8 Seed Stability 26

C When Does Correction Fail? A EuroSAT Case Study **26**

D Additional Comparison and Transfer Results **28**

D.1 Domain Generalization 28

D.2 Detailed Comparison with NeRP 28

D.3 Inference Efficiency 29

E Extension to Autoregressive Inference **30**

F Geometric Interpretation **31**

F.1 Classifier Boundary Along the REP-Induced Direction 31

F.2 First-Cross Rule as Minimum Intervention 32

F.3 Frozen Semantic Verification 32

F.4 Summary of the Correction Mechanism 33

G Broader Related Work and Positioning **33**

G.1 Base-to-Novel Adaptation and Representation Specialization 33

G.2 Plug-in Enhancement, Calibration, and Prediction-Stage Correction 34

G.3 Internal Representations as Inference-Time Signals 34

A METHOD AND IMPLEMENTATION DETAILS

This appendix provides implementation details and a complete inference algorithm, followed by additional diagnostics and controlled analyses that complement Sec. 3.2. We then analyze the EuroSAT failure case, report detailed comparison and transfer results, evaluate an autoregressive extension, derive the geometry of the bounded first-cross correction, and conclude with broader positioning relative to adaptation and inference-time intervention methods.

A.1 HOST-SPECIFIC REP READOUT

As described in the main paper, REPURPOSE does not introduce additional trainable parameters or modify the original adaptation procedure. Instead, it obtains an adaptation-specific representation through a fixed, parameter-free readout from the trained host.

For an input image x, we denote the normalized CLS representation by $z_{\mathrm{CLS}}(x)$ and the corresponding adaptation-specific representation by

$$z_{\mathrm{REP}}(x) \in \mathbb{R}^d. \tag{10}$$

REP refers to hidden states produced after processing the current image, rather than to the learnable input embeddings themselves. Therefore, $z_{\mathrm{REP}}(x)$ is sample-dependent.

Although the underlying token constructions differ across adaptation methods, we use a common readout principle whenever possible: the relevant final-layer states are aggregated, mapped through the CLS-aligned visual projection, and ℓ_2-normalized. This places z_{CLS} and z_{REP} in the same visual embedding space while preserving their distinct functional roles.

MMRL and MMRL++. For MMRL and MMRL++, REP is obtained from the dedicated representation tokens introduced by the corresponding host. Let $\{h_{\mathrm{REP}}^{(i)}\}_{i=1}^{N_r}$ denote their final hidden states. We first aggregate these states as

$$\bar{h}_{\mathrm{REP}} = \frac{1}{N_r} \sum_{i=1}^{N_r} h_{\mathrm{REP}}^{(i)}, \tag{11}$$

and then apply the host visual projection followed by ℓ_2 normalization:

$$z_{\mathrm{REP}} = \mathrm{norm}\left(P_{\mathrm{vis}} \bar{h}_{\mathrm{REP}}\right). \tag{12}$$

MaPLe. For MaPLe, REP is read from the final-layer visual prompt hidden states. Let $\{h_{\mathrm{prompt}}^{(i)}\}_{i=1}^{N_p}$ denote these states. We mean-pool the prompt states and map the result through the CLS-aligned visual projection:

$$z_{\mathrm{REP}} = \mathrm{norm}\left(P_{\mathrm{vis}} \frac{1}{N_p} \sum_{i=1}^{N_p} h_{\mathrm{prompt}}^{(i)}\right). \tag{13}$$

PromptSRC. For PromptSRC, we analogously read the final visual prompt states, mean-pool them, apply the corresponding visual projection, and ℓ_2-normalize the resulting feature:

$$z_{\mathrm{REP}} = \mathrm{norm}\left(P_{\mathrm{vis}} \frac{1}{N_p} \sum_{i=1}^{N_p} h_{\mathrm{prompt}}^{(i)}\right). \tag{14}$$

These host-specific states are not assumed to be equivalent or to serve the same purpose in their original formulations. REP provides only a common interface for studying whether adaptation-specific information already present inside a trained host can serve as corrective evidence during novel-class inference.

A.2 IMPLEMENTATION CONFIGURATION

Table 5 summarizes the shared inference settings and the host-specific REP configurations used throughout our experiments. All evaluations use the released host checkpoints without retraining, fine-tuning, or parameter updates.

Table 5: **Implementation configuration.** The CLIP backbone and REPURPOSE settings are fixed across datasets. Host-specific rows describe the adaptation states used to instantiate REP.

Item	Scope	Configuration
CLIP backbone	All hosts	ViT-B/16; frozen at inference
Evaluation seeds	All hosts	1, 2, 3
Checkpoint use	All hosts	Official released checkpoints; no retraining or test-time parameter updates
Semantic prototypes	Default	Released CoPrompt descriptions; standard class-name semantics are used only in controlled analyses
Global semantic energy	All hosts	$\rho = 0.95$ for selecting the global semantic rank r_g
Candidate set	All hosts	Entropy-adaptive $K(x)$ from Eq. 3, clipped to $[2, C]$
Correction rule	All hosts	Bounded first crossing along $z(\gamma)$ with $0 < \gamma \leq 1$
Verifier	All hosts	Frozen-CLIP semantic preference $F_{c \to j^\star}(x)$
B2N safeguard	Base-to-novel	$\tau_{\text{ver}} = 0.07$ in Eq. 38
Transfer verification	CD / DG	Original zero-threshold rule $F_{c \to j^\star}(x) > 0$; no base–novel partition is defined
MMRL / MMRL++ REP	MMRL, MMRL++	5 representation tokens; REP pathway configured over visual layers 6–12; final REP states are mean-pooled
MaPLe REP	MaPLe	Final-layer visual prompt states; $N_{\text{CTX}} = 2$ and prompt depth 9
PromptSRC REP	PromptSRC	Final visual prompt states; $4 + 4$ context-token configuration and prompt depth 9

The host-specific token configurations above affect only how the already trained adaptation state is read out. Once $z_{\text{REP}}(x)$ is obtained, the same semantic projection, bounded first-crossing rule, and verification logic are used for all hosts.

A.3 Semantic Prototype Construction

We construct frozen semantic prototypes from the released CoPrompt descriptions. No additional prompt template is applied to these descriptions.

For ImageNet, Caltech101, DTD, Food101, StanfordCars, SUN397, UCF101, and FGVCAircraft, we use 50 descriptions per class. For EuroSAT, Flowers102, and OxfordPets, we use 10 descriptions per class.

For class c, let

$$\mathcal{D}_c = \{d_{c,m}\}_{m=1}^{M_c} \tag{15}$$

denote its textual descriptions. Each description is directly encoded by the frozen CLIP text encoder E_{text} and individually ℓ_2-normalized:

$$e_{c,m} = \text{norm}\left(E_{\text{text}}(d_{c,m})\right). \tag{16}$$

The class-level semantic prototype is obtained by averaging the normalized description embeddings and normalizing the resulting mean:

$$p_c = \text{norm}\left(\frac{1}{M_c}\sum_{m=1}^{M_c} e_{c,m}\right). \tag{17}$$

Thus, normalization is applied both before and after class-level aggregation. The standard CLIP class-name prompt is used only in the semantic-resource ablation in Appendix B.5.

A.4 Global Semantic Basis

Let

$$P = \begin{bmatrix} p_1^\top \\ \vdots \\ p_C^\top \end{bmatrix} \in \mathbb{R}^{C \times d} \tag{18}$$

collect the frozen semantic prototypes of all C classes.

We first compute the mean semantic prototype

$$\bar{p} = \frac{1}{C}\sum_{c=1}^{C} p_c \tag{19}$$

and center the prototype matrix:

$$\widetilde{P} = P - \mathbf{1}\bar{p}^\top. \tag{20}$$

We then compute the compact singular value decomposition

$$\widetilde{P} = U\Sigma V^\top, \tag{21}$$

implemented using `torch.linalg.svd(full_matrices=False)`.

Let σ_i denote the singular values in descending order. Given a retained semantic-energy ratio ρ, we choose the smallest global rank r_g satisfying

$$\frac{\sum_{i=1}^{r_g} \sigma_i^2}{\sum_i \sigma_i^2} \geq \rho. \tag{22}$$

Unless otherwise specified, we use $\rho = 0.95$.

The resulting global semantic basis is

$$B_g = V_{[:,1:r_g]} \in \mathbb{R}^{d \times r_g}. \tag{23}$$

A.5 Instance-Adaptive Local Semantic Basis

The global basis captures dataset-level semantic variation, whereas the local basis focuses on semantic directions relevant to the classes that are plausible for the current sample.

Let $\pi = \mathrm{softmax}(\ell)$ denote the host predictive distribution, following the notation in the main text. We compute its entropy as

$$H(x) = -\sum_{c \in \mathcal{Y}} \pi_c \log \pi_c. \tag{24}$$

The instance-adaptive candidate-set size is

$$K(x) = \min\left(C, \max\left(2, \left\lceil e^{H(x)} \right\rceil\right)\right). \tag{25}$$

We denote by $\mathcal{C}_K(x)$ the top-$K(x)$ classes under the original host classifier.

Let

$$P_{\mathcal{C}_K(x)} \in \mathbb{R}^{K(x) \times d} \tag{26}$$

collect the frozen semantic prototypes associated with these candidate classes. We first center the candidate prototypes by their own mean:

$$\bar{p}_{\mathcal{C}_K(x)} = \frac{1}{K(x)} \sum_{c \in \mathcal{C}_K(x)} p_c, \tag{27}$$

$$\widetilde{P}_{\mathcal{C}_K(x)} = P_{\mathcal{C}_K(x)} - \mathbf{1}\bar{p}_{\mathcal{C}_K(x)}^\top. \tag{28}$$

We then project the centered candidate structure into the global semantic coordinates:

$$Q(x) = \widetilde{P}_{\mathcal{C}_K(x)} B_g \in \mathbb{R}^{K(x) \times r_g}. \tag{29}$$

A compact SVD is performed in these global semantic coordinates:

$$Q(x) = U_\ell \Sigma_\ell V_\ell^\top. \tag{30}$$

Unlike the global construction, the local stage does not apply an additional semantic-energy threshold. Instead, its retained rank is

$$r_\ell(x) = \min\left(K(x) - 1, r_g\right). \tag{31}$$

The $K(x) - 1$ term reflects the maximum rank of the centered candidate prototype matrix.

Let

$$V_{\ell,r} = V_{\ell[:,1:r_\ell(x)]}. \tag{32}$$

The corresponding local directions in the original embedding space are obtained by lifting these directions through the global basis:

$$B(x) = B_g V_{\ell,r} \in \mathbb{R}^{d \times r_\ell(x)}. \tag{33}$$

Finally, the candidate-relevant component of REP is

$$\delta(x) = B(x)B(x)^\top z_{\mathrm{REP}}(x). \tag{34}$$

Importantly, the retained-energy criterion is used only to determine the global rank r_g. The local construction instead uses candidate centering, projection into the global semantic subspace, local SVD, and the rank constraint in Eq. 31; no additional local energy truncation is applied.

A.6 Inference Procedure

Given the original CLS representation $z_{\mathrm{CLS}}(x)$ and the candidate-relevant REP component $\delta(x)$, REPurpose uses the same bounded one-dimensional correction path as in Eq. 5:

$$z(\gamma) = z_{\mathrm{CLS}}(x) + \gamma\delta(x), \qquad 0 < \gamma \leq 1. \tag{35}$$

The original host representation corresponds to the starting point $\gamma = 0$. The upper bound prevents extrapolation beyond the extracted REP evidence.

The adapted classifier remains unchanged. Rather than freely combining CLS and REP predictions, REPurpose identifies the first classifier boundary crossed within this bounded interval along the REP-induced direction. The corresponding class transition is treated only as a proposal and is independently checked using frozen CLIP semantics. If the verifier rejects the proposal, the original host prediction is retained.

This design gives REP a deliberately restricted role: it supplies corrective evidence but never serves as an independent novel-class classifier. No host parameters, classifier weights, or prompts are updated during inference, and REPurpose performs no test-time optimization.

Algorithm 1 summarizes the complete inference procedure.

Algorithm 1 REPurpose inference.

Require: Adapted host: z_{CLS}, z_{REP}, adapted text features $\{t_c^a\}$ and logits ℓ; frozen semantic prototypes $\{p_c\}$ with global basis (95% energy); frozen image feature $v(x)$; threshold ϵ_D (default 0; safeguard $\epsilon_D = \tau_{\mathrm{ver}}\,[\log(r_h/r_f)]_+$).
Ensure: Corrected prediction $\hat{y}$.

$c \leftarrow \arg\max_c \ell_c$ {host top-1}
$H(x) \leftarrow -\sum_c \pi_c \log \pi_c$ for $\pi = \mathrm{softmax}(\ell)$; $K(x) \leftarrow \min\{C, \max\{2, \lceil e^{H(x)} \rceil\}\}$; $\mathcal{C}_K(x) \leftarrow$ top-$K(x)$ classes equation 3.
$B(x) \leftarrow$ orthonormal basis of centered $\{p_j\}_{j \in \mathcal{C}_K(x)}$ within the global basis (Appendix A.5).
$\delta(x) \leftarrow B(x)B(x)^\top z_{\mathrm{REP}}$ equation 4.
$\mathcal{R}(x) \leftarrow \emptyset$
for $j \in \mathcal{C}_K(x) \setminus \{c\}$ **do**
  $\gamma_j \leftarrow \dfrac{z_{\mathrm{CLS}}^\top (t_c^a - t_j^a)}{\delta(x)^\top (t_j^a - t_c^a)}$ equation 6
  if $\delta(x)^\top (t_j^a - t_c^a) > 0$ **and** $0 < \gamma_j \leq 1$ **then**
    $\mathcal{R}(x) \leftarrow \mathcal{R}(x) \cup \{j\}$
  end if
end for
if $\mathcal{R}(x) \neq \emptyset$ **then**
  $j^\star \leftarrow \arg\min_{j \in \mathcal{R}(x)} \gamma_j$ equation 7
  if $v(x)^\top (p_{j^\star} - p_c) > \epsilon_D$ equation 8 **then**
    $\hat{y} \leftarrow j^\star$ {accept $c \to j^\star$}
  else
    $\hat{y} \leftarrow c$ {reject}
  end if
else
  $\hat{y} \leftarrow c$ {no reachable candidate}
end if
return $\hat{y}$

Semantic prototypes and the global semantic basis are computed once per dataset and reused across samples. For base-to-novel evaluation, ϵ_D is computed once per dataset and seed from the label-free response statistics r_h and r_f, and the resulting threshold is reused for all samples in that evaluation unit. Only the sample-dependent steps in Algorithm 1 are evaluated for each input.

B Additional Analyses

Unless noted otherwise, these analyses use the official MMRL checkpoints, visual REP, and the standard 11-dataset base-to-novel protocol. Results are averaged over three seeds.

B.1 Additional Motivation Diagnostics

Per-dataset linear probing. Table 6 expands the diagnostic linear-probe results from Sec. 3.2. For each host, probes are trained on the novel-class 16-shot training split (the standard per-seed B2N splits, with 16 images per novel class) and evaluated on the full novel test split; no test samples are used for probe training or hyperparameter selection. CLS and REP features are extracted from the frozen host and row-wise ℓ_2-normalized, and the concatenated setting uses $z = [z_{\mathrm{CLS}}; z_{\mathrm{REP}}] \in \mathbb{R}^{2d}$. Each probe is a single linear layer with bias, optimized using cross-entropy with ℓ_2 regularization and L-BFGS (learning rate 1.0, at most 200 iterations, strong-Wolfe line search, and history size 20). The regularization coefficient $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ is selected by 5-fold stratified cross-validation using only the 16-shot training split. For the shuffled control, REP features are randomly permuted across samples within each split, with train and test features shuffled independently while avoiding same-class swaps; the permutation is fixed for each checkpoint–split seed pair. This preserves the REP distribution and dimensionality while breaking the image–REP correspondence. Results for each dataset are averaged over three official checkpoint seeds and five 16-shot split seeds (15 evaluation units in total), and the reported macro result is the unweighted mean over the 11 datasets. Novel-class labels are used solely for this diagnostic analysis and never enter REPurpose inference.

Table 6: **Per-dataset linear-probe accuracy on novel classes.** Shuff. denotes CLS concatenated with REP features shuffled across samples. The macro row is the unweighted mean over the 11 datasets.

	MMRL				MMRL++			
Dataset	CLS	REP	CLS+REP	Shuff.	CLS	REP	CLS+REP	Shuff.
Caltech101	96.87	96.64	97.13	95.49	96.78	96.35	96.85	95.23
DTD	79.99	79.53	81.05	73.45	80.35	79.34	81.43	72.39
EuroSAT	95.57	94.41	95.55	94.54	96.04	96.86	97.18	92.79
FGVCAircraft	63.92	65.11	67.09	54.61	65.85	65.72	67.48	53.95
Food101	90.14	88.54	90.29	85.61	89.95	89.17	90.13	85.26
ImageNet	71.05	70.26	72.17	63.36	70.91	70.48	71.93	63.86
Flowers102	98.47	98.16	98.58	96.09	98.51	98.17	98.56	95.79
OxfordPets	96.90	96.72	97.31	95.17	96.85	96.99	97.20	93.81
StanfordCars	89.10	89.03	90.62	77.68	89.24	87.67	89.60	77.77
SUN397	83.46	82.08	84.09	75.62	83.65	80.66	83.27	75.69
UCF101	88.99	86.01	89.28	84.16	89.11	86.21	88.95	83.81
Macro	**86.77**	**86.05**	**87.56**	**81.44**	**87.02**	**86.15**	**87.51**	**80.94**

On MMRL, matched CLS+REP improves the macro accuracy by 0.79 points over CLS and exceeds CLS on 10 of the 11 datasets; EuroSAT is the only exception (-0.02 points). This is the setting underlying the 10/11 statement in the main text. On MMRL++, the macro gain is 0.48 points and CLS+REP exceeds CLS on 9 of 11 datasets, with small decreases on SUN397 and UCF101. In both hosts, shuffling REP causes a large drop, supporting the conclusion that the useful REP information is sample-specific rather than a consequence of increased feature dimensionality alone.

Cross-host frozen-CLIP corrective potential. The main text reports that standard class-name zero-shot frozen CLIP correctly recognizes 6,851 of the 43,100 novel-class errors made by MMRL. We repeat the same error-conditioned diagnostic for all four hosts and additionally report the corresponding result when frozen semantics are formed from the CoPrompt descriptions used by the default REPURPOSE pipeline.

Table 7: **Frozen semantic recoverability on host errors.** Counts are pooled over 11 datasets and three seeds. "Correct" counts host errors for which the frozen semantic classifier predicts the ground-truth novel class; percentages use the host-error count in the second column as the denominator.

Host	Host errors	Class-name correct	CoPrompt correct
MMRL	43,100	6,851 (15.9%)	8,945 (20.8%)
MMRL++	41,736	6,513 (15.6%)	8,182 (19.6%)
MaPLe	45,214	9,919 (21.9%)	11,337 (25.1%)
PromptSRC	44,752	7,171 (16.0%)	10,379 (23.2%)

The class-name MMRL row exactly matches the 15.9% statistic used in the main text. The same phenomenon is present for every host: a non-trivial subset of host errors remains correctly recognizable by frozen CLIP. CoPrompt descriptions increase this recoverable subset for all four hosts, while the class-name column provides the directly matched zero-shot diagnostic used in Sec. 3.2. These numbers are error-conditioned pooled statistics and should not be confused with the 11-dataset macro accuracies in the main result tables.

B.2 Detailed Analysis of Rank-Local Host Errors

We provide detailed statistics for the rank-locality analysis in Sec. 3.2. All statistics are computed on misclassified novel-class samples. Table 8 reports results pooled over all error samples from 11 datasets and three seeds, matching the aggregation used in the main text.

Table 8: **Cross-host rank-locality statistics on novel-class errors.** Results are pooled over all misclassified samples from 11 datasets and three seeds. GT@k denotes the percentage of errors for which the ground-truth class remains within the host top-k predictions. ECR denotes the coverage of the entropy-adaptive candidate set, and K/C denotes the fraction of the label space retained by the candidate set. All quantities in this table, including Avg. K, are computed only over misclassified novel samples and pooled over errors.

Host	GT@2	GT@3	GT@5	ECR	Avg. K/C	Avg. K	Median GT Rank	#Errors
MMRL	47.39	63.99	77.77	80.52	7.99	7.45	3.0	43,100
MMRL++	46.34	63.54	77.34	79.36	6.62	7.13	3.0	41,736
MaPLe	46.81	63.11	77.40	79.71	8.97	7.57	3.0	45,214
PromptSRC	47.14	63.75	77.58	81.94	9.50	8.74	3.0	44,752

For completeness, Table 9 reports the complementary dataset-macro view of the same error-conditioned analysis. Macro-averaging assigns equal weight to datasets with different label-space sizes and numbers of errors, and therefore yields larger average K/C values than pooling all errors together.

Across hosts, the dataset-level median ground-truth rank remains approximately 2.7–3.0. Under pooled accounting, a fixed top-5 set occupies 9.25% of the label space for MMRL and 7.22% for MMRL++, providing a useful reference for the compactness of the adaptive candidate set. For MMRL, error-conditioned Avg. K is 5.23 under dataset-macro aggregation and 7.45 when all errors are pooled.

These values use a different population from the candidate-set ablation, where Avg. K is computed over all novel test samples within each dataset before dataset-level macro-averaging (3.55 for MMRL). The difference is expected: since $K(x)$ is determined by the entropy-adaptive rule in Eq. 3,

Table 9: **Dataset-macro rank-locality statistics on novel-class errors.** Each metric is first computed within a dataset over its misclassified novel samples and then macro-averaged over the 11 datasets. This aggregation differs from the pooled statistics in Table 8.

Host	GT@5	ECR	Avg. K/C
MMRL	81.73	81.78	14.83
MMRL++	82.25	81.69	14.75
MaPLe	81.71	79.43	14.10
PromptSRC	80.47	81.11	15.40

misclassified samples tend to have higher predictive entropy and therefore receive larger candidate sets.

B.3 Alternative Inference Controls

Table 10 gives the full fusion sweep behind Table 2(b). Frozen-CLIP reranking uses the same entropy-adaptive candidate set as REPURPOSE. The fusion controls add either raw REP or its semantic projection directly to CLS.

Table 10: **Semantic reranking and direct-fusion controls on MMRL.**

Variant	Novel Acc.
CLS only	77.17
Frozen-CLIP rerank	77.78
$\text{CLS} + 0.25\,\text{REP}$	77.26
$\text{CLS} + 0.50\,\text{REP}$	76.93
$\text{CLS} + 1.00\,\text{REP}$	76.13
$\text{CLS} + 0.25\,\text{Proj}(\text{REP})$	77.70
$\text{CLS} + 0.50\,\text{Proj}(\text{REP})$	77.66
$\text{CLS} + 1.00\,\text{Proj}(\text{REP})$	77.31
REPURPOSE	**78.30**

Raw REP becomes harmful as its fusion weight increases. Projection makes direct fusion safer, but it still falls short of REPURPOSE. The same ordering holds on MMRL++: REPURPOSE reaches 79.25%, compared with 77.97% for frozen-CLIP reranking and 78.86% for the best fusion control.

B.4 Candidate-Set Size

We compare the entropy-adaptive candidate set with fixed values of K. GT coverage is measured on host errors and denotes the fraction for which the ground-truth class is included in the candidate set.

Table 11: **Candidate-set size and coverage.** Avg. K is first averaged over *all* novel test samples within each dataset and then macro-averaged across the 11 datasets, whereas GT coverage is computed only over misclassified novel samples.

Rule	Avg. K	GT coverage (%)	Novel Acc.
$K = 2$	2.00	54.98	78.26
$K = 3$	3.00	70.39	78.26
$K = 5$	5.00	81.73	78.28
$K = 10$	10.00	91.63	78.16
Adaptive	3.55	**81.78**	**78.30**

The adaptive rule reaches essentially the same error coverage as $K = 5$ (81.78% versus 81.73%) while using 3.55 candidates on average, 29% fewer than a fixed set of five. Their novel accuracies are also effectively tied (78.30% versus 78.28%). Thus, the main benefit of the adaptive rule is

a more compact candidate set at comparable coverage, rather than a distinct accuracy gain over $K = 5$.

B.5 SEMANTIC RESOURCE ABLATION

Our default semantic prototypes are constructed from the publicly released CoPrompt descriptions. Since richer descriptions may themselves improve correction, we perform controlled ablations to disentangle their contribution from that of REP. We first conduct a 2×2 factorial analysis over the correction evidence (semantic-only vs. REP) and semantic resource (standard CLIP class-name prompts vs. CoPrompt descriptions), and then isolate where the richer semantics contribute within REPURPOSE.

REP and semantic resources. The semantic-only control removes REP evidence while retaining the same candidate construction and semantic verification pipeline. All four settings use the same final verification protocol, and results are averaged over the 11 datasets and three seeds.

Table 12: **Factorial analysis of REP and semantic resources.** Novel accuracy (%) averaged over 11 datasets.

Host	Evidence	Class-name	CoPrompt
MMRL	Semantic-only	75.15	77.78
	+ REP	77.37	**78.30**
MMRL++	Semantic-only	75.66	77.97
	+ REP	78.30	**79.25**

CoPrompt descriptions provide a clear semantic benefit. Without REP, replacing class-name semantics with CoPrompt improves MMRL from 75.15% to 77.78% (+2.63 points) and MMRL++ from 75.66% to 77.97% (+2.31 points). Thus, richer semantic descriptions are an important contributor to the strongest configuration.

Importantly, however, this semantic advantage does not account for the contribution of REP. Holding CoPrompt semantics fixed, adding REP improves MMRL from 77.78% to 78.30% (+0.52 points) and MMRL++ from 77.97% to 79.25% (+1.29 points). The same pattern holds, and is even stronger, with standard class-name semantics: adding REP improves MMRL from 75.15% to 77.37% (+2.22 points) and MMRL++ from 75.66% to 78.30% (+2.64 points). REP therefore provides additional corrective value under both semantic resources rather than relying on the richer CoPrompt descriptions.

The two factors are not additive. The factorial interaction is -1.71 points for MMRL and -1.35 points for MMRL++, indicating partially overlapping benefits between REP and richer semantics. We therefore avoid attributing a fixed fraction of the final improvement to either factor; instead, the matched comparisons above show that both contribute under controlled settings.

Where do richer semantics contribute? Semantic prototypes enter REPURPOSE at two stages: they define the semantic subspace used to project REP, and they are used by the frozen semantic verifier. To localize the contribution of richer descriptions, we vary the semantic resource independently at these two stages.

Table 13: **Location-specific semantic-resource ablation.** Rows specify the semantic resource used for REP projection and columns the resource used for semantic verification. Results are novel accuracy (%) averaged over 11 datasets.

Host	Projection	Class-name	CoPrompt
MMRL	Class-name	77.371	77.889
	CoPrompt	77.415	**78.298**
MMRL++	Class-name	78.295	78.799
	CoPrompt	78.373	**79.253**

The isolated effect of richer descriptions is substantially larger at verification than at REP projection. With class-name projection fixed, replacing only the verifier semantics with CoPrompt improves MMRL by 0.52 points (77.371%→77.889%) and MMRL++ by 0.50 points (78.295%→78.799%). In contrast, with the class-name verifier fixed, replacing only the projection semantics yields smaller gains of 0.04 points for MMRL (77.371%→77.415%) and 0.08 points for MMRL++ (78.295%→78.373%). Using CoPrompt at both stages produces the strongest results, reaching 78.298% and 79.253%, respectively. These results suggest that richer descriptions primarily strengthen semantic verification, while the REP projection itself is relatively insensitive to semantic-resource richness. The additional gain when CoPrompt is used at both stages also indicates a positive interaction between projection and verification semantics.

Interaction with the safeguard on EuroSAT. We finally examine how the semantic resource interacts with the response-based safeguard introduced for the MMRL++/EuroSAT failure case. This interaction is relevant because changing the semantic prototypes also changes the frozen verifier's novel-block response rate r_f, and can therefore change whether the safeguard is activated.

For MMRL with CoPrompt semantics, the safeguard is inactive and the final EuroSAT accuracy is 80.61%. For MMRL++, CoPrompt semantics trigger the safeguard, which recovers the original zero-threshold result from 86.31% to 88.23%. These results illustrate that safeguard activation is determined by the response mismatch rather than by dataset identity alone.

More generally, changing from CoPrompt descriptions to class-name semantics can alter r_f and hence the safeguard activation pattern. We therefore use the same final verification protocol for all cells in the factorial and location-specific analyses above, rather than selectively enabling the safeguard for a particular semantic resource. This ensures that the reported comparisons reflect the behavior of each semantic resource under the same inference rule.

B.6 Semantic Energy Sensitivity

We vary the retained semantic-energy ratio ρ used to set the global semantic rank.

Table 14: **Sensitivity to retained semantic energy on MMRL.**

Retained energy ρ	Novel Acc.
0.90	78.114
0.95	78.296
0.99	**78.446**
1.00	78.376

Performance is relatively stable across this range. Although $\rho = 0.99$ is numerically highest in this sweep, we keep $\rho = 0.95$ as a single fixed default across datasets and hosts rather than selecting the best value per dataset.

B.7 Crossing-Order Analysis

We examine how corrective information is distributed along the REP-induced trajectory. For each sample, we order all reachable decision boundaries by their crossing distance γ and evaluate proposals from successive crossings under the same frozen verifier.

Table 15: **Correction quality across successive reachable crossings.** Results aggregate MMRL and MMRL++ over 11 datasets and three seeds under the same verifier.

Crossing	Macro Acc.	W→C	C→W	W→W	Prop. Rate	Precision
First	**78.69**	**7,567**	4,557	7,016	4.81%	**51.89%**
Second	77.91	1,108	585	3,347	1.27%	24.87%
Third	77.78	259	162	1,537	0.49%	14.42%
Last reachable	78.39	6,331	4,229	6,316	4.24%	50.06%

The first crossing provides the strongest corrective signal, achieving the highest macro accuracy and correction precision. Correction precision drops from 51.89% at the first crossing to 24.87%

and 14.42% at the second and third crossings, respectively. Consistently, the ground-truth hit rate along the trajectory decreases from 14.98% to 2.33% and 0.60%. These results indicate that useful corrective information is strongly concentrated at the earliest decision transition.

B.8 SEED STABILITY

Table 16 summarizes paired gains across the three host seeds. The intervals are 95% t-confidence intervals over the three paired seed differences.

Table 16: **Seed stability of Novel-accuracy gains.** We report paired gains over three seeds, where each seed's gain is computed between REPURPOSE and the corresponding host using the 11-dataset macro Novel accuracy.

Host	Mean gain $\pm$ SD (pp)	95% CI (pp)
MaPLe	$+1.60 \pm 0.20$	$[1.11, 2.09]$
PromptSRC	$+1.32 \pm 0.12$	$[1.03, 1.61]$
MMRL	$+1.13 \pm 0.32$	$[0.33, 1.93]$
MMRL++	$+0.93 \pm 0.13$	$[0.60, 1.26]$

For MaPLe and PromptSRC, the paired gains are computed against our reproduced host results (75.08% and 76.09% Novel accuracy, respectively), while the main comparison table retains the originally reported host results. For MMRL and MMRL++, the reproduced host results match the reported values to two decimal places.

All four paired mean gains are positive, and all four confidence intervals exclude zero. With only three seeds, we report these intervals as a compact measure of seed-to-seed variation rather than as a large-sample estimate.

C WHEN DOES CORRECTION FAIL? A EUROSAT CASE STUDY

The MMRL++ result on EuroSAT provides a useful case for understanding when correction can become harmful. We analyze the same official checkpoints, visual REP readout, CoPrompt semantic prototypes, and three seeds used in the main evaluation. In particular, we examine whether the failure comes from the REP correction signal itself or from the frozen semantic verification.

Table 17: **Failure diagnosis on EuroSAT.** Results use official checkpoints and are averaged over three seeds. Proposal rate is the fraction of samples for which the bounded first-cross rule produces a candidate transition. Verifier margin is the median frozen semantic preference among accepted transitions. Transition counts are summed over the three seeds.

Host	Host Acc.	Corrected Acc.	Proposal	Verifier margin	$W \to C$ / $C \to W$
MMRL	80.15	80.61	5.8%	0.0778	181 / 128
MMRL++	88.22	86.31	10.0%	0.0342	119 / 343

Weak verification under a strong adapted host. EuroSAT presents a large gap between the adapted host and frozen CLIP. MMRL++ reaches 88.22% Novel accuracy, whereas frozen CLIP reaches only 64.05%. Although frozen CLIP is used only to verify a proposed transition rather than to make the prediction itself, this gap makes its semantic preference a weaker verification signal for the adapted host.

As shown in Table 17, MMRL++ proposes transitions for 10.0% of samples, compared with 5.8% for MMRL, while its median accepted verification margin is only 0.0342, less than half the MMRL value of 0.0778. Under the default zero-threshold rule, MMRL++ produces 343 correct-to-wrong transitions but only 119 wrong-to-correct transitions, reducing accuracy from 88.22% to 86.31%. MMRL shows the opposite behavior, with 181 wrong-to-correct versus 128 correct-to-wrong transitions and an improvement from 80.15% to 80.61%. This suggests that the MMRL++/EuroSAT regression mainly comes from insufficient verification selectivity.

The REP signal remains useful. The regression does not mean that REP loses its corrective information. On host errors, the ground-truth class is contained in the candidate set for 89.3% of samples. Among proposed transitions, the first-cross target matches the ground truth for 32.0% with the corresponding REP, compared with 11.4% after shuffling REP across samples. In addition, the ground-truth boundary is reachable along the bounded REP trajectory for 77.8% of candidate-covered errors. These results suggest that useful corrective information remains in REP, while the main failure occurs when deciding whether to accept a proposal.

Verification threshold. We first test whether stronger verification can address this failure by replacing $F_{c\to j^\star}(x) > 0$ with

$$F_{c\to j^\star}(x) > \epsilon. \tag{36}$$

For MMRL++/EuroSAT, $\epsilon = 0.01$ and 0.02 improve accuracy from 86.31% to 88.07% and 88.25%, respectively. However, applying the same thresholds to MMRL/EuroSAT reduces its corrected accuracy from 80.61% to 80.42% and 80.30%. A fixed positive threshold is therefore useful for the MMRL++ failure case but is not appropriate for every host.

We instead increase the threshold only when the adapted host shows a stronger novel-class response than frozen CLIP. Let r_h and r_f denote the fractions of novel-test samples predicted into the novel-class block by the adapted host and frozen CLIP, respectively, using the shared base+novel label space. We use the label-free mismatch

$$I_{\text{gap}} = \left[\log \frac{r_h}{r_f}\right]_+, \tag{37}$$

and set the verification threshold to

$$\epsilon_D = \tau_{\text{ver}} I_{\text{gap}}. \tag{38}$$

Here r_h and r_f are prediction frequencies and do not use ground-truth labels. When $r_h \le r_f$, $I_{\text{gap}} = 0$ and the original zero-threshold verifier is recovered.

We fix $\tau_{\text{ver}} = 0.07$ for all datasets and seeds, following the temperature initialization used in CLIP (Radford et al., 2021). For the three MMRL++/EuroSAT seeds, I_{gap} is 0.373, 0.355, and 0.189, giving thresholds of 0.0261, 0.0249, and 0.0132, respectively. For MMRL/EuroSAT, $r_h < r_f$, so the safeguard is inactive and its original beneficial correction is preserved. With this adjustment, MMRL++/EuroSAT improves from 86.31% to 88.23%, compared with 88.22% for the host.

Table 18 makes the activation pattern explicit. Across the $2 \times 11 \times 3 = 66$ MMRL/MMRL++ host–dataset–seed units, the only material activations occur on the three MMRL++/EuroSAT runs. Two OxfordPets runs exhibit only negligible activation ($\epsilon_D < 0.0007$) and do not affect the reported result; the remaining units are effectively at the zero-threshold rule.

Table 18: **Material safeguard activations.** The table lists the three MMRL++/EuroSAT runs for which the response mismatch meaningfully increases the verification threshold. With $\tau_{\text{ver}} = 0.07$, $\epsilon_D = \tau_{\text{ver}} I_{\text{gap}}$.

Unit	r_h	r_f	I_{gap}	ϵ_D
MMRL++ / EuroSAT / seed 1	0.792	0.545	0.373	0.0261
MMRL++ / EuroSAT / seed 2	0.778	0.545	0.355	0.0249
MMRL++ / EuroSAT / seed 3	0.659	0.545	0.189	0.0132

Sensitivity. Table 19 shows that the result is insensitive to the exact value of τ_{ver}.

Over the range $[0.05, 0.15]$, MMRL++/EuroSAT varies only from 88.18% to 88.26%, while its 11-dataset average remains between 79.25% and 79.26%. MMRL remains unchanged at 78.30% throughout. We therefore use the fixed value 0.07 without dataset-specific tuning.

Semantic-resource control. The original regression is also not explained by the use of richer CoPrompt descriptions. Under the original zero-threshold verifier, replacing them with standard CLIP class-name prompts improves MMRL++/EuroSAT from 86.31% to 86.99%, but this remains below the 88.22% host. For MMRL, the same class-name control reaches 81.57%. Thus, changing the semantic resource alone does not resolve the failure.

Table 19: **Sensitivity to τ_{ver}.** EuroSAT reports MMRL++ Novel accuracy, while the last two rows report 11-dataset Novel averages.

τ_{ver}	0.03	0.05	0.07	0.08	0.09	0.10	0.15	0.20
MMRL++ / EuroSAT	87.97	88.18	88.23	88.26	88.26	88.24	88.22	88.21
MMRL++ / Avg.	79.23	79.25	79.25	79.26	79.26	79.25	79.25	79.25
MMRL / Avg.	78.30	78.30	78.30	78.30	78.30	78.30	78.30	78.30

D Additional Comparison and Transfer Results

We provide additional results complementing the comparisons reported in the main paper, including detailed domain-generalization results, the matched NeRP reproduction, and inference-efficiency analysis.

D.1 Domain Generalization

We further evaluate ImageNet-adapted models on four ImageNet distribution shifts.

Table 20: **Detailed domain-generalization results.** Accuracy (%) on four ImageNet distribution shifts. Avg. denotes the macro-average over the four targets.

Method	ImageNet-V2	ImageNet-Sketch	ImageNet-A	ImageNet-R	Avg.
MMRL	64.52	49.32	51.03	77.21	60.52
MMRL + REPURPOSE	64.87	50.08	51.63	78.17	**61.19**
MMRL++	64.31	49.15	50.74	77.47	60.42
MMRL++ + REPURPOSE	64.81	50.07	51.10	78.24	**61.06**

For MMRL, the average accuracy increases from 60.52% to 61.19%; for MMRL++, it increases from 60.42% to 61.06%. Both hosts improve on all four evaluated distribution shifts.

D.2 Detailed Comparison with NeRP

We provide the complete NeRP reproduction under the matched MMRL setting. For base-to-novel (B2N) evaluation, we use the official MMRL `train_base` checkpoints. For cross-dataset (CD) transfer and domain generalization (DG), NeRP and the MMRL reference use the same P2 ImageNet CD checkpoint from `cd_dg/train`. We use NeRP's released inference configuration without retuning it on target-test accuracy.

Unless otherwise specified, NeRP results are reported as the mean and standard deviation over three seeds. We additionally report the corresponding MMRL CLS accuracy as a matched-host reference.

Base-to-novel generalization. Table 21 reports novel-class accuracy on all 11 base-to-novel benchmarks.

Table 21: **Per-dataset base-to-novel comparison with NeRP.** Novel-class accuracy (%) is reported on the 11 B2N datasets. NeRP results are mean±std over three seeds. MMRL CLS is included as the matched-host reference. Macro denotes the average across datasets.

Method	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	Macro
MMRL CLS	71.43	94.47	97.58	75.09	77.26	91.52	37.05	79.29	64.94	80.15	80.06	77.17
MMRL + NeRP	71.57±0.12	94.77±0.12	97.77±0.32	75.13±0.55	78.77±0.23	91.57±0.12	37.47±0.21	79.53±0.21	65.83±1.01	78.03±2.90	80.77±0.57	77.38

NeRP increases the B2N macro-average from 77.17% for the matched MMRL CLS reference to 77.38%. The per-dataset results also reveal substantial variation in the effect of the correction, with EuroSAT showing particularly large seed sensitivity. We examine this case separately below.

Cross-dataset transfer. For cross-dataset transfer, the model is adapted on ImageNet and evaluated directly on ten downstream datasets without further training. Both NeRP and the MMRL reference use the same ImageNet CD checkpoint.

Table 22: **Per-dataset cross-dataset comparison with NeRP.** Accuracy (%) is reported on ten target datasets after adaptation on ImageNet. NeRP results are mean±std over three seeds. MMRL CLS uses the same ImageNet CD checkpoint.

Method	Caltech101	OxfordPets	StanfordCars	Food101	DTD	UCF101	Flowers102	FGVCAircraft	SUN397	EuroSAT	Macro
MMRL CLS	94.56	91.37	65.78	86.35	46.36	68.75	72.47	26.38	67.55	53.01	67.26
MMRL + NeRP	94.77±0.21	91.40±0.20	66.13±0.42	86.33±0.06	46.73±1.00	69.80±0.10	73.20±0.70	26.27±0.49	67.63±0.23	55.47±1.97	67.77

On cross-dataset transfer, NeRP increases the macro-average from 67.26% to 67.77% under the matched MMRL checkpoint.

Domain generalization. For domain generalization, we evaluate the same ImageNet-adapted checkpoint on ImageNet-V2, ImageNet-Sketch, ImageNet-A, and ImageNet-R.

Table 23: **Per-domain comparison with NeRP.** Accuracy (%) is reported on four ImageNet distribution shifts. NeRP results are mean±std over three seeds.

Method	ImageNet-V2	ImageNet-Sketch	ImageNet-A	ImageNet-R	Macro
MMRL CLS	64.52	49.32	51.03	77.21	60.52
MMRL + NeRP	65.40±0.10	48.80±0.26	51.77±0.23	77.33±0.42	60.83

NeRP increases the DG macro-average from 60.52% to 60.83%. Its effect is not uniform across shifts: accuracy increases on ImageNet-V2, ImageNet-A, and ImageNet-R, while decreasing on ImageNet-Sketch.

Aggregate comparison with REPURPOSE. For completeness, Table 24 summarizes the aggregate comparison under the matched MMRL setting.

Table 24: **Aggregate comparison under the MMRL host.** B2N denotes base-to-novel generalization, CD denotes cross-dataset transfer, and DG denotes domain generalization. All values are accuracies (%).

Method	B2N	CD	DG
MMRL CLS	77.17	67.26	60.52
MMRL + NeRP	77.38	67.77	60.83
MMRL + REPURPOSE	**78.30**	**68.14**	**61.19**

Under the matched MMRL setting, NeRP reaches 77.38%, 67.77%, and 60.83% on B2N, CD, and DG, respectively. With the default frozen verifier, REPURPOSE reaches 78.30%, 68.14%, and 61.19% on the corresponding evaluation settings.

Seed sensitivity on EuroSAT. EuroSAT exhibits substantially larger seed variation in the NeRP reproduction than most other B2N datasets. We therefore report the individual seed results in Table 25.

The direction of NeRP's effect on EuroSAT changes across host seeds: it decreases accuracy for Seeds 1 and 2 but increases accuracy for Seed 3. Correspondingly, the three-seed NeRP result on EuroSAT has a standard deviation of 2.90 percentage points. We therefore report the complete seed-level reproduction rather than characterizing this dataset using a single checkpoint.

D.3 INFERENCE EFFICIENCY

We additionally compare inference efficiency under a matched evaluation protocol. Measurements use the same GPU, MMRL checkpoint, batch size, worker count, warm-up procedure, and timing loop within each software environment. Runtime is measured on an NVIDIA 4090 GPU using the ImageNet `new/test` split containing approximately 25K images, with batch size 64, four data-loading workers, 20 warm-up iterations, and two timed runs.

Because NeRP and REPURPOSE are implemented in different software environments, absolute runtimes across the two environments can include environment-specific overhead. We therefore normalize each method by the bare MMRL runtime measured in its corresponding environment.

Table 25: **Seed-level B2N results on EuroSAT.** Accuracy (%) is reported for the matched MMRL host and NeRP reproduction. The final row reports the three-seed mean.

Seed	MMRL CLS	MMRL + NeRP
1	85.05	75.50
2	82.33	77.40
3	73.08	81.20
Mean	80.15	78.03

Table 26: **Inference efficiency under the MMRL host.** Runtime is reported in milliseconds per image. Relative runtime is normalized by the bare MMRL runtime measured in the same software environment.

Method	ms/image	Relative runtime
MMRL (NeRP env.)	1.784	1.00×
MMRL + NeRP	3.013	1.69×
MMRL (ours env.)	1.730	1.00×
MMRL + REPURPOSE (adapted verifier)	2.020	1.17×
MMRL + REPURPOSE (frozen verifier)	2.333	1.35×

The frozen-verifier configuration incurs a $1.35\times$ relative runtime compared with its bare MMRL baseline, while the adapted-verifier configuration requires $1.17\times$. Under its corresponding matched environment, NeRP requires $1.69\times$ the bare MMRL runtime. Neither NeRP nor REPURPOSE performs test-time optimization.

E Extension to Autoregressive Inference

We further evaluate whether REPURPOSE can reuse adaptation-specific representations in an autoregressive vision–language model. We use M2PT (Wang et al., 2024b) on CIFAR-10 and CIFAR-100, with text and visual prompt lengths of 10 and 20, respectively. The M2PT checkpoint contains 20 visual prompt tokens at each of the 24 visual Transformer layers; we explicitly restore all layer-wise visual prompts from the checkpoint.

Visual representation readout. For an input image, the final visual encoder output contains 20 visual prompt tokens, one CLS token, and 256 image patch tokens. Let h_{CLS} denote the final CLS state and $\{h_i^{\mathrm{vp}}\}_{i=1}^{20}$ the final visual prompt states. We obtain the CLS and REP representations as

$$z_{\mathrm{CLS}} = \mathrm{norm}(W_v\ \mathrm{LN}(h_{\mathrm{CLS}}))\,, \tag{39}$$

$$z_{\mathrm{REP}} = \mathrm{norm}\left(W_v\ \mathrm{LN}\left(\frac{1}{20}\sum_{i=1}^{20} h_i^{\mathrm{vp}}\right)\right), \tag{40}$$

where LN and W_v are the CLIP visual post-layer normalization and projection. Thus, the visual prompt states provide REP, while the CLS state provides the reference representation for correction.

REP-guided correction. We construct frozen CLIP prototypes p_c using the template "a photo of a {class name}." and obtain the visual host decision

$$c = \arg\max_k z_{\mathrm{CLS}}^{\top} p_k. \tag{41}$$

The remaining correction procedure follows the main method. We construct the global semantic basis using the 95% energy criterion, select an adaptive candidate set from the CLS prediction, form the candidate-local semantic basis, and project z_{REP} onto this basis. The resulting direction is used to find the first reachable class boundary within the bounded path $0 < \gamma \leq 1$. If no boundary is reachable, the original decision is kept.

For a proposed transition $c \to j^\star$, we use the frozen CLIP image representation $v(x)$ to compute

$$F_{c\to j^\star}(x) = v(x)^{\top}(p_{j^\star} - p_c). \tag{42}$$

The transition is accepted only when $F_{c\to j^\star}(x) > 0$. We use the frozen verifier in this experiment because the CLS decision and the adapted verifier would use the same representation and prototypes: by definition of c, $z_{\mathrm{CLS}}^\top(p_j - p_c) \le 0$ for every $j \neq c$. An adapted verifier would therefore reject every proposed transition.

Autoregressive output correction. REPURPOSE does not intervene in M2PT decoding. We first run the original M2PT generation procedure to obtain a response y_{raw}. If REPURPOSE accepts a transition to $j^\star$, the final response is the corresponding class name; otherwise, the original M2PT response is retained:

$$y_{\mathrm{final}} = \begin{cases} \mathrm{name}(j^\star), & \text{if the correction is accepted,} \\ y_{\mathrm{raw}}, & \text{otherwise.} \end{cases} \tag{43}$$

No decoder hidden states, token logits, KV cache, sampling probabilities, or stopping criteria are modified. This experiment therefore evaluates REPURPOSE as a representation-guided post-hoc correction rather than a decoder-time intervention.

Evaluation protocol. The M2PT evaluation files and cached visual representations use different sample orders. We align them through the original CIFAR image indices and verify complete one-to-one coverage and label consistency before applying the correction decisions. Both the original and corrected generations are evaluated with the same Vicuna-13B-v1.5 judging protocol.

Under this matched protocol, M2PT obtains 89.28% on CIFAR-10 and 59.21% on CIFAR-100. With REPURPOSE, the corresponding accuracies are 89.31% and 60.62%, respectively.

F GEOMETRIC INTERPRETATION

We provide a geometric interpretation of the correction mechanism, focusing on the bounded first-cross rule and its minimum-intervention property along the REP-induced correction direction.

F.1 CLASSIFIER BOUNDARY ALONG THE REP-INDUCED DIRECTION

Let the original host prediction be

$$c = R_0(x) = \arg\max_{c\in\mathcal{Y}} z_{\mathrm{CLS}}^\top t_c^a, \tag{44}$$

where t_c^a denotes the adapted text representation for class c. Let $\mathcal{C}_K(x)$ denote the candidate set defined in the main text, which contains the original prediction c.

For clarity, consider the unnormalized bounded correction path

$$z(\gamma) = z_{\mathrm{CLS}} + \gamma\delta(x), \qquad 0 \le \gamma \le 1, \tag{45}$$

where $\delta(x)$ is the REP-induced correction direction for the current sample.

At any fixed γ, normalizing $z(\gamma)$ multiplies all class scores by the same positive scalar and therefore does not change their ordering or the location of pairwise classifier boundaries along the path. Hence, boundary crossings can be analyzed directly using $z(\gamma)$.

Consider a candidate class $j \in \mathcal{C}_K(x) \setminus \{c\}$. A pairwise boundary crossing between c and j occurs when

$$z(\gamma)^\top t_j^a = z(\gamma)^\top t_c^a. \tag{46}$$

Substituting Eq. 45 gives

$$(z_{\mathrm{CLS}} + \gamma\delta(x))^\top \left(t_j^a - t_c^a\right) = 0. \tag{47}$$

Because c is the original host prediction, its initial pairwise margin over j satisfies

$$m_j := z_{\mathrm{CLS}}^\top \left(t_c^a - t_j^a\right) \ge 0. \tag{48}$$

The rate at which this margin is reduced along the correction direction is

$$s_j := \delta(x)^\top \left(t_j^a - t_c^a\right). \tag{49}$$

If $s_j \leq 0$, moving along $\delta(x)$ cannot close the initial margin, and therefore no positive crossing with class j is reachable along the correction path. If $s_j > 0$, the pairwise crossing coefficient is

$$\gamma_j = \frac{m_j}{s_j} = \frac{z_{\mathrm{CLS}}^\top \left(t_c^a - t_j^a\right)}{\delta(x)^\top \left(t_j^a - t_c^a\right)}. \tag{50}$$

Because the correction path is bounded to $0 \leq \gamma \leq 1$, a candidate transition is feasible only when

$$0 < \gamma_j \leq 1. \tag{51}$$

Thus, a candidate whose boundary lies beyond $\gamma = 1$ is not regarded as reachable, even if its score would eventually overtake that of c under an unbounded extrapolation along $\delta(x)$.

We therefore define the set of feasible candidate-class crossings as

$$\mathcal{R}(x) = \{j \in \mathcal{C}_K(x) \setminus \{c\} \ : \ s_j > 0,\ 0 < \gamma_j \leq 1\}. \tag{52}$$

F.2 First-Cross Rule as Minimum Intervention

If $\mathcal{R}(x)$ is nonempty, the first-cross rule selects

$$j^\star = \arg\min_{j \in \mathcal{R}(x)} \gamma_j. \tag{53}$$

If $\mathcal{R}(x)$ is empty, no candidate-class boundary is reachable within the bounded correction path and the original prediction c is retained.

The first-cross rule admits a simple minimum-intervention interpretation. Along the fixed REP-induced direction $\delta(x)$, the pre-normalization displacement from the original CLS representation is

$$\Delta z(\gamma) = z(\gamma) - z_{\mathrm{CLS}} = \gamma \delta(x), \tag{54}$$

and hence

$$\|\Delta z(\gamma)\|_2 = \gamma \|\delta(x)\|_2, \qquad 0 \leq \gamma \leq 1. \tag{55}$$

For a fixed sample, $\delta(x)$ is fixed and therefore $\|\delta(x)\|_2$ is constant. Since every feasible crossing satisfies $\gamma_j > 0$, it follows that

$$\arg\min_{j \in \mathcal{R}(x)} \gamma_j = \arg\min_{j \in \mathcal{R}(x)} \|\Delta z(\gamma_j)\|_2. \tag{56}$$

Thus, among the candidate-class boundaries reachable within the bounded REP-induced path, the first-cross rule selects the transition requiring the smallest pre-normalization displacement from the original CLS representation along $\delta(x)$.

This is a directional and candidate-restricted minimum-intervention property. It does not claim that the selected transition is the globally minimum-distance perturbation over arbitrary directions or over unrestricted classes in the full embedding space. Instead, REPurpose deliberately restricts the admissible correction in three ways: the alternative class must belong to the local candidate set, the correction must follow the REP-induced direction, and the corresponding boundary must be reached within the bounded interval $0 < \gamma \leq 1$. Within these constraints, the first-cross rule selects the earliest feasible decision boundary.

F.3 Frozen Semantic Verification

The first-cross rule identifies the candidate-class transition requiring the smallest movement along the bounded REP-induced direction. However, geometric reachability alone does not imply that the proposed transition is semantically preferable.

Let $v(x)$ denote the frozen CLIP visual representation of the image and p_c the frozen semantic prototype of class c. For the proposed transition $c \to j^\star$, we compute the frozen semantic preference

$$F_{c \to j^\star}(x) = v(x)^\top \left(p_{j^\star} - p_c\right). \tag{57}$$

The proposed transition is accepted only when

$$F_{c\to j^\star}(x) > \epsilon_D. \tag{58}$$

Otherwise, the original prediction c is retained. The default rule uses $\epsilon_D = 0$; in base-to-novel evaluation, the label-free safeguard in Eq. 38 may increase this threshold when the adapted host and frozen CLIP exhibit a novel-response mismatch.

Equivalently, Eq. 58 requires

$$v(x)^\top p_{j^\star} > v(x)^\top p_c + \epsilon_D. \tag{59}$$

Importantly, the frozen verifier is applied only to the single transition proposed by the adapted host and the REP-induced correction direction. It does not independently search over the label space and therefore does not replace the original adapted classifier. Instead, it acts only as an independent semantic check on the locally proposed decision change.

F.4 Summary of the Correction Mechanism

The complete correction can be interpreted as a sequence of constrained decisions.

First, the host prediction defines a local candidate set, and semantic projection restricts REP to directions supported by the candidate-relevant frozen semantic subspace. Second, the bounded first-cross rule considers only candidate-class boundaries reachable within $0 < \gamma \leq 1$ and selects the earliest such boundary along the REP-induced direction. Third, frozen semantic verification accepts the proposed transition only when pretrained CLIP semantics independently prefer the target class over the original prediction.

Together, these constraints allow adaptation-specific information to influence novel-class inference without treating REP as a standalone predictor, modifying the trained host, extrapolating arbitrarily far along the REP-induced direction, or allowing frozen CLIP semantics to freely replace the original CLS decision.

G Broader Related Work and Positioning

REPurpose lies at the intersection of vision–language model adaptation, prediction-stage correction, and inference-time use of latent model signals. The closest works differ in an important aspect of the problem they address: most adaptation methods determine *what representations should be learned*, while most inference-time correction methods introduce a new corrective signal from pretrained predictions, reference inputs, or manipulated activations. Our focus is instead on whether an adaptation-specific representation that is already produced by a trained host, but is unreliable as a standalone novel-class predictor, can still serve as useful corrective evidence for the host's transferable prediction.

G.1 Base-to-Novel Adaptation and Representation Specialization

A central problem in adapting pretrained vision–language models is to acquire task-specific information without destroying the general knowledge inherited from pretraining. Early prompt-learning methods such as CoOp and CoCoOp adapt CLIP through learnable textual context, while MaPLe extends prompt learning to both visual and textual branches and couples the two modalities during adaptation Zhou et al. (2022b;a); Khattak et al. (2023a). These methods established prompt learning as an effective parameter-efficient mechanism for downstream adaptation, but also exposed the tension between specialization to base classes and transfer to novel classes.

A number of subsequent approaches explicitly address this specialization– generalization trade-off. PromptSRC regularizes prompted representations toward the frozen pretrained model and uses prompt ensembling and textual diversity to preserve generalizable information Khattak et al. (2023b). TaskRes separates the frozen pretrained classifier from a learnable task-specific residual, thereby making the distinction between prior and task-specific knowledge explicit Yu et al. (2023). DePT similarly attributes the base-to-new trade-off to the interference between base-specific and task-shared information, and moves base-specific knowledge into a decoupled feature space. These

methods primarily intervene during *adaptation*: they seek to learn a representation or classifier whose organization itself better preserves transferability. Zhang et al. (2024)

More recent methods make the functional separation between transferable and specialized representations even more explicit. MMRL introduces dedicated representation tokens for dataset- and task-related information while preserving a more transferable class pathway Guo & Gu (2025). Its inference strategy reflects this distinction by using specialized representation features for base recognition while relying on the more generalizable class features for new tasks. MMRL++ further develops this representation-learning paradigm with parameter-efficient and interaction-aware representation learning Guo & Gu (2026). Related approaches, including DePT and AlignedNorm, likewise study how adaptation-induced structure can be separated, preserved, or aligned to improve novel-class generalization Zhang et al. (2024); Ma et al. (2026).

REPurpose is complementary to this line of work. It does not propose a new training objective for separating task-specific and transferable information. Instead, it starts from a host in which specialization has already occurred and asks whether the resulting specialized states have been discarded too aggressively at novel-class inference. In particular, our results motivate a distinction between *predictive reliability* and *corrective informativeness*: a representation may be poorly aligned with the fixed novel-class decision rule and therefore unsuitable as an independent predictor, while still containing sample-specific evidence that can help resolve a nearby ambiguity in the transferable representation.

G.2 Plug-in Enhancement, Calibration, and Prediction-Stage Correction

A related body of work improves an already pretrained or adapted VLM through auxiliary guidance or plug-in mechanisms rather than replacing the host model. Some approaches modify the adaptation process by introducing additional semantic supervision, pretrained guidance, or complementary representation structure. These methods demonstrate that knowledge sources beyond the standard prompted classifier can improve the balance between specialization and generalization. Their additional signals, however, are typically injected during training or are explicitly learned as part of the adaptation procedure.

Closer to our setting are methods that operate at prediction time. Distance- Aware Calibration (DAC) studies the open-vocabulary calibration problem of fine-tuned CLIP and adjusts prediction confidence according to the relationship between predicted labels and base classes Wang et al. (2024a). Thus, DAC modifies the behavior of an adapted classifier after training, but its correction is based on calibration structure in the label space rather than an adaptation-specific representation produced for the current image.

NeRP is particularly relevant because it also performs local correction for novel-class inference without modifying the trained host Tian et al. (2026). It identifies asymmetric pretrained inter-class preferences using neutral textual prompts and reference images, combines these priors with sample-level evidence, and conditionally flips locally ambiguous predictions. REPurpose shares with NeRP the principle that novel-class errors are often local and can be addressed conservatively rather than through global reclassification. The source of corrective evidence is nevertheless different. NeRP estimates a pretrained inter-class preference from neutral references, whereas REPurpose asks whether information already encoded in the *adapted host's own specialized representation* can provide the sample-specific correction direction.

This distinction also determines the role assigned to the corrective signal. REP in our framework is not converted into a second classifier and is not freely ensembled with the original prediction. It defines only a candidate-relevant direction in representation space. The unchanged adapted classifier determines whether a bounded transition is geometrically reachable and which admissible boundary is encountered first, after which frozen pretrained semantics independently verify the proposed transition. Thus, the host remains the decision geometry rather than one member of an unconstrained ensemble.

G.3 Internal Representations as Inference-Time Signals

The idea that a model's internal computation may contain useful information not fully reflected in its final output has broad precedent outside CLIP adaptation. In language models, Inference-

Time Intervention (ITI) identifies activation directions associated with truthful behavior and modifies selected internal activations during inference Li et al. (2023b). DoLa instead exploits differences across transformer depth: it projects earlier and later layers to the vocabulary space and contrasts their logits during decoding, based on the observation that useful factual information can emerge differently across layers Chuang et al. (2024). These works show that information exposed by intermediate model states can be useful even when it is not directly expressed by the default output distribution.

Similar ideas have appeared in large multimodal models. Visual Contrastive Decoding (VCD) modifies generation by contrasting predictions obtained from original and distorted visual inputs, reducing reliance on language priors without additional model training Leng et al. (2024). Although VCD operates on contrasted output distributions rather than a latent representation from a single forward pass, it illustrates the broader usefulness of conservative inference-time correction in multimodal models. Jiang et al. analyze intermediate attention patterns in large vision–language models and show that middle-layer visual processing exposes signals associated with object hallucination; they use these signals to perform a training-free attention adjustment at inference time Jiang et al. (2025). SteerVLM more directly uses latent multimodal embeddings to learn lightweight activation modulation for controlling VLM behavior at inference time Sivakumar et al. (2025).

These studies provide important precedent for treating latent model states as actionable inference-time signals. REPurpose differs in both the origin and the intended semantics of the representation. We do not search arbitrary hidden layers for a behavioral steering direction, nor do we learn an intervention vector for a desired property. REP is a representation with an explicit origin in downstream adaptation: it is produced by the same trained host for the current image and encodes information introduced through its specialization process. Our question is therefore narrower: when such a representation is not sufficiently transferable to serve as a novel-class predictor, can it nevertheless provide evidence for correcting the host's transferable prediction?

Table 27: Positioning of REPurpose relative to representative adaptation, prediction-correction, and inference-intervention methods. "Internal signal" indicates that the corrective signal is obtained from latent states produced inside the evaluated model.

Method	Primary setting	Correction / auxiliary signal	Internal signal	Role at inference	Additional correction training
TaskRes	CLIP adaptation	Task-specific classifier residual	No	Adapted classifier	Yes
DePT	CLIP adaptation	Decoupled base-specific features	Yes	Learned representation	Yes
MMRL / MMRL++	CLIP adaptation	Specialized representation tokens	Yes	Host representation	Yes
DAC	Adapted CLIP	Label–base-class distance	No	Calibration	No
NeRP	Adapted CLIP	Pretrained preference + sample evidence	No	Local prediction correction	No
ITI	LLMs	Learned activation directions	Yes	Activation intervention	Lightweight
DoLa	LLMs	Early/late-layer logits	Yes	Contrastive decoding	No
VCD	LVLMs	Original/distorted-input predictions	No	Contrastive decoding	No
Devils in Middle Layers	LVLMs	Intermediate attention patterns	Yes	Attention correction	No
SteerVLM	LVLMs	Latent multimodal activations	Yes	Activation steering	Lightweight
REPurpose	**Adapted CLIP**	**Adaptation-specific REP**	**Yes**	**Bounded corrective evidence**	**No**