AUDITING VLM REPAIR: VARIANCE GEOMETRY AND GUARDED CANDIDATE SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Repairing a frozen vision-language model (VLM) can improve average accuracy while damaging its worst-performing groups. We formulate VLM repair as a deployment-time audit and selection problem: repair methods generate candidates, and a separate policy chooses which candidate to apply. Our main group-free policy selects by validation-average accuracy within a pool of cosine and explicit nuisance-direction repairs. We also audit candidate risk using variance geometry, validation-average preservation, and prompt stability. Across 156 scenarios spanning Waterbirds, CelebA, synthetic CIFAR shifts, PACS, and two VLM backbones, naive validation-average selection yields a worst-group/domain delta of -0.006, whereas the family-guarded policy yields $+0.037$ with a 95% bootstrap interval of $[+0.003, +0.078]$ and reduces harmful selections from 30 to 9. The combined risk signal identifies harmful observed-covariance candidates with AUROC 0.956. In a separate fixed-gate stress test on 120 shared Waterbirds and CelebA scenarios, adding TIE-style candidates yields a robust delta of $+0.073$ and a paired gain of $+0.041$ over the base pool under the same gate. These results establish candidate selection as a consequential layer of VLM repair: it improves robust deployment outcomes and enables specialized repair methods to strengthen an audited pool.

1 INTRODUCTION

Vision-language models (VLMs) transfer a shared representation from image and language supervision to downstream recognition without task-specific training (Radford et al., 2021). This interface is attractive precisely because the model can be adapted after pre-training. A practitioner can change prompt templates, transform the embedding geometry, remove a suspected concept, or average predictions over augmented views while keeping the encoders frozen. Recent methods have made this design space unusually productive: prompt learning and test-time prompting improve transfer (Zhou et al., 2022b;a; Shu et al., 2022; Hassan et al., 2024; Yoon et al., 2024), training-free cache adaptation provides another frozen-encoder route (Zhang et al., 2022), label-free prompt distributions and bias correction improve zero-shot accuracy (Zhu et al., 2024b), and fine-tuning-free debiasing targets spurious attributes directly (Dehdashtian et al., 2024; Gerych et al., 2024).

The deployment question is different from the method-development question. A repair method usually returns one preferred transformation after optimizing a validation objective. In a shifted environment, however, the high-variance directions of the embedding space can represent class semantics, a dataset-specific background, or an irrelevant acquisition artifact. A covariance-aware distance therefore has no fixed interpretation. Whitening can amplify a useful direction, leave a text prototype unchanged, or magnify a nuisance direction. The failure is especially consequential under group shift, where average accuracy hides a small group with a large error rate (Sagawa et al., 2020). Domain-generalization benchmarks such as PACS make the same issue visible across visual styles (Li et al., 2017a), while WILDS broadens the evaluation view to naturally occurring deployment shifts (Koh et al., 2021).

This paper asks a deployment-level question: *given a pool of plausible VLM repairs, which candidate should be applied when group labels are unavailable?* Its accept/reject structure is related to selective prediction with a reject option (Geifman & El-Yaniv, 2019), but our decision acts on repair candidates rather than individual examples. Representation, prompt, and test-time methods

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.

supply possible repairs; the selection problem concerns their robust behavior after deployment. This distinction matters across adaptation mechanisms: TTA changes predictions and calibration (Farina et al., 2024; Yoon et al., 2024), concept erasure removes a linear attribute (Belrose et al., 2023), and spectral steering exploits a text subspace (Dafnis & Metaxas, 2025). Their value as candidates does not determine which repair a deployment should choose.

We develop an audit framework with distinct selection and diagnostic roles. The main policy, GUARDED-REPAIR, restricts deployment to cosine and text- or pseudo-label-based nuisance-direction repairs, then selects by validation-average accuracy with deterministic tie-breaks. This family guard needs no group labels. Separately, a candidate-risk audit combines variance alignment, validation-average preservation, and prompt stability. Its fixed signal gate is used in the external-candidate stress test. Keeping these roles explicit lets us evaluate deployment outcomes, diagnose harmful candidates, and test the value of an enlarged repair pool.

Figure 1 summarizes the evidence. The best covariance-aware candidates improve visual-prototype classification, while the text zero-shot CIFAR-100 control selects the cosine endpoint. Across the main shift suite, family-guarded selection changes a negative robust delta into a positive one and reduces harmful deployments. The risk panel establishes the value of combining audit signals: geometry alone has AUROC 0.293 on observed-covariance candidates, while the combined signal reaches 0.956.

Our contributions are:

- We formulate VLM repair as an auditable candidate-selection problem and develop a group-free family guard that separates repair generation from deployment choice.
- We establish task-conditioned covariance behavior and evaluate complementary risk signals, reaching AUROC 0.956 on harmful observed-covariance candidates.
- Across 156 scenarios, the main policy changes robust delta from -0.6 to $+3.7$ points and reduces harmful selections from 30 to 9. An external TIE-style pool reaches $+7.3$ robust points under the fixed signal gate.

The central contribution is the deployment decision: candidate-level auditing makes the benefits and risks of VLM repair explicit. Specialized methods such as TIE-style translation strengthen this framework by supplying additional repair candidates.

2 RELATED WORK

Vision-language transfer and test-time adaptation. CLIP established zero-shot classification by aligning image and text representations at scale (Radford et al., 2021), while SigLIP changed the contrastive objective and provides a second widely used VLM family (Zhai et al., 2023). Prompt learning adapts the language side with few labels (Zhou et al., 2022b;a), and cache-based adapters provide a training-free alternative (Zhang et al., 2022). More recent test-time methods use prompt tuning (Shu et al., 2022), distribution alignment (Hassan et al., 2024), text-feature dispersion for calibration (Yoon et al., 2024), label-free prompt distributions (Zhu et al., 2024b), entropy-minimization adaptation (Wang et al., 2021), or image augmentation and voting without backpropagation (Farina et al., 2024). Spectrum-aware steering explicitly constructs a semantic subspace for latent adaptation (Dafnis & Metaxas, 2025). These methods motivate our candidate pool. Our focus is the decision made after candidates are generated, with worst-group/domain robustness as a first-class deployment metric.

Debiasing and concept removal. FairerCLIP formulates joint representation debiasing in RKHSs and supports labeled and pseudo-labeled settings (Dehdashtian et al., 2024). BendVLM uses input-dependent nonlinear transformations (Gerych et al., 2024), while our external translation-style candidate moves image embeddings along a spurious text direction. LEACE gives a closed-form linear concept-erasure operator (Belrose et al., 2023). These approaches target the repair itself. We use their mechanisms as candidates or external comparisons, then test whether a group-free audit can identify when the repair should be trusted. This separation matters because the same concept-removal operator can improve Waterbirds while reducing CelebA robustness in our controlled feature-cache comparison.

Robustness under group and domain shift. Worst-group metrics expose failures that average accuracy hides, and group DRO established a standard language for such shifts (Sagawa et al., 2020). PACS provides a complementary domain-shift benchmark with photo, art, cartoon, and sketch domains (Li et al., 2017a). The domain-generalization literature also shows that model selection is itself a major source of variation (Gulrajani & Lopez-Paz, 2021); our evaluation applies that concern to frozen VLM repair candidates rather than training algorithms. WILDS further motivates reporting deployment shifts beyond i.i.d. accuracy (Koh et al., 2021). We also include the recent variance-reduction fine-tuning line (Zhu et al., 2024a) as a conceptual neighbor: both studies treat variance as consequential, but our setting is training-free candidate selection and our output is a deployment decision.

3 Problem Setting and Audit Objective

Let an image encoder and a text encoder map an image x and class description c to normalized embeddings $f(x) \in \mathbb{R}^d$ and $p_c \in \mathbb{R}^d$. The cosine baseline predicts

$$\hat{y}(x) = \arg\max_c f(x)^\top p_c. \tag{1}$$

We assume access to an unlabeled or class-labeled fit subset, a validation subset with class labels, and an unlabeled test stream. Deployable fitting and selection exclude group labels; benchmark construction, reporting, and oracle comparisons are described in Section 5. A *candidate repair* is a deterministic score function h_m constructed from the fit subset and a prompt set. The full audit pool includes cosine, covariance-aware transformations, nuisance-direction filters, and external repair candidates.

For each scenario s, we measure average accuracy A_s and a robust metric R_s. The robust metric is worst-group accuracy for Waterbirds, CelebA, and synthetic CIFAR shifts, and worst-domain accuracy for PACS. We report deltas relative to the cosine baseline:

$$\Delta A_s = A_s(h_m) - A_s(h_{\cos}), \qquad \Delta R_s = R_s(h_m) - R_s(h_{\cos}). \tag{2}$$

The deployment problem is to choose m using only fit and validation information. We evaluate a policy by its aggregate ΔR and its number of harmful selections, defined as scenarios with $\Delta R_s < 0$.

4 Task-Conditioned Variance Audit

4.1 Covariance operators and candidate repairs

Given fit embeddings $z_1, \ldots, z_n$, we compute the centered observed covariance

$$\widehat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \bar{z})(z_i - \bar{z})^\top. \tag{3}$$

To control small eigenvalues, we use shrinkage

$$\widehat{\Sigma}_\alpha = (1-\alpha)\widehat{\Sigma} + \alpha \bar{\lambda} I + 10^{-6} \bar{\lambda} I, \tag{4}$$

where $\bar{\lambda} = \operatorname{tr}(\widehat{\Sigma})/d$ and α is selected on validation average within the candidate pool. The corresponding whitened score is cosine similarity after multiplication by $W_\alpha = \widehat{\Sigma}_\alpha^{-1/2}$. The isotropic endpoint $\alpha = 1$ is exactly the cosine geometry up to a common scale.

We also construct a top-variance filter. Let U_k contain the first k eigenvectors of $\widehat{\Sigma}$, and define $P_k = I - U_k U_k^\top$. The filtered score is

$$s_k(x, c) = \cos(P_k f(x), P_k p_c). \tag{5}$$

For a binary nuisance attribute, a text nuisance direction is the difference between two nuisance prompt prototypes. A pseudo nuisance direction replaces the attribute labels with predictions from the nuisance prompt prototypes and computes within-class mean differences. For PACS, we use centered class and domain prototype sets and retain the corresponding low-rank direction matrices.

Algorithm 1 Main policy: group-free family-guarded selection.

Require: Fit embeddings, class-labeled validation set, candidate pool $\mathcal{H}$

1: Form the eligible pool from cosine and text/pseudo nuisance-direction candidates.
2: Include the corresponding text/pseudo domain-direction candidates for PACS.
3: Evaluate validation-average accuracy for each eligible candidate.
4: Return the candidate with highest validation average; break ties by complexity, parameter, and deterministic method name.

4.2 Variance alignment index

Let $\widehat{\Sigma} = U \operatorname{diag}(\lambda_1, \ldots, \lambda_d) U^\top$ with eigenvalues sorted in descending order. Given semantic directions D_s and nuisance directions D_n, we define energy in eigendirection j as

$$e_j^s = \frac{1}{|D_s|} \sum_{v \in D_s} (v^\top u_j)^2, \qquad e_j^n = \frac{1}{|D_n|} \sum_{v \in D_n} (v^\top u_j)^2. \tag{6}$$

The semantic-to-nuisance ratio is $q_j = (e_j^s + \epsilon)/(e_j^n + \epsilon)$. Our variance-alignment index over the leading $K = 128$ directions is

$$\mathrm{VAI}_{128} = \mathrm{Spearman}\left(\log(\lambda_j + \epsilon), \log(q_j)\right)_{j=1}^{128}. \tag{7}$$

Positive alignment means that high-variance directions are relatively more semantic than nuisance. The index is a diagnostic, not a standalone decision rule. In particular, a low value can describe a failure boundary but cannot by itself identify every harmful candidate.

4.3 Group-free family-guarded selection

The main policy, GUARDED-REPAIR, admits cosine and explicit text- or pseudo-label-based nuisance-direction candidates. PACS uses their domain-direction counterparts. Observed-covariance and top-variance operators remain in the full audit pool but are excluded from this family-guarded deployment pool. Among admitted candidates, selection maximizes validation-average accuracy. Ties favor lower operator complexity, then lower parameter value, with a deterministic method-name tie-break. The policy uses no validation worst-group accuracy, group labels, or test outcomes. The group-informed oracle instead selects by validation worst-group/domain accuracy and serves as a reference for the remaining selection gap.

4.4 Candidate-risk diagnostics and fixed signal gate

The signal audit evaluates individual candidates separately from the family guard. Validation-average preservation requires $\Delta A_{\mathrm{val}} \geq -0.005$. Prompt stability requires that at least 0.60 of the predefined class-prompt sets pass that threshold, holding the other candidate settings fixed. Observed-covariance candidates additionally require $\mathrm{VAI}_{128} \geq 0.05$; explicit nuisance-direction candidates bypass this geometry check. The fixed signal gate accepts candidates passing all applicable checks and then selects by validation average with cosine retained as fallback. TIE-style candidates use the validation and prompt checks. This gate governs the external-pool experiment, not the main policy in Algorithm 1.

For the observed-covariance risk analysis, let b_A, b_P, and b_V indicate failure of the validation, prompt, and geometry checks. The combined risk score is $2b_A + b_P + 2b_V$. AUROC evaluates this score as a ranking of harmful candidates; the fixed accept/reject decision rejects any applicable check failure. Candidate diagnosis and deployed-policy accuracy are evaluated separately.

5 Experimental Protocol

5.1 Benchmarks and backbones

The clean geometry track uses visual class prototypes on CIFAR-100 (Krizhevsky, 2009) with CLIP ViT-B/32, DINOv2 ViT-B/14 (Oquab et al., 2023), and MAE ViT-B/16 (He et al., 2022), plus PathMNIST from MedMNIST (Yang et al., 2023) with DINOv2. WebVision-1k (Li et al., 2017b) measures

Table 1: Covariance-aware gains relative to cosine, in percentage points. Visual-prototype rows use the best covariance-aware candidate over three seeds; WebVision uses observed whitening with a paired bootstrap over one feature cache.

Setting	Prototype / backbone	Gain
CIFAR-100	visual / CLIP	$+8.3$
CIFAR-100	visual / DINOv2	$+3.2$
CIFAR-100	visual / MAE	$+25.1$
PathMNIST	visual / DINOv2	$+9.0$
WebVision-1k	visual / CLIP	$+0.7$
CIFAR-100	text zero-shot / CLIP	$+0.0$

a noisy-web training distribution against a clean validation set with CLIP. True text zero-shot controls use CLIP text prototypes on CIFAR-100 and Waterbirds.

The main audit suite contains 156 scenarios. Waterbirds (Sagawa et al., 2020) contributes 30 scenarios from 3 class-only fit-subset seeds, 5 prompt sets, and two backbones (CLIP ViT-B/32 and SigLIP ViT-B/16). CelebA Blond-Hair x Male (Liu et al., 2015) contributes 90 scenarios from 3 correlations, 3 fit seeds, 5 prompt sets, and two backbones. Synthetic CIFAR border and tint shifts contribute 9 scenarios each across three strengths. PACS contributes 18 scenarios from 3 class-domain-stratified split seeds, 3 prompt sets, and two backbones. All feature extraction and model inference remain on the server-side repository; the paper directory contains only source files and figures.

5.2 Protocol and uncertainty

Waterbirds deployable fit subsets are stratified by class only. Place labels are excluded from subset construction and candidate selection. CelebA deployable nuisance directions use CLIP or SigLIP nuisance prompts and pseudo labels; true Male labels appear only in oracle rows and evaluation. PACS domain labels define the benchmark split and robust metric, but are excluded from deployable fitting. These controls isolate the value of the audit from direct access to nuisance annotations.

Main-policy confidence intervals use 5,000 bootstrap replicates. Within each dataset/backbone, clusters fix the prompt set and correlation, variant, and strength where applicable, averaging repeated seeds; PACS clusters fix the prompt set. Dataset-level combined-backbone rows resample backbone means. The overall row averages available backbones within each dataset and resamples equally weighted dataset means. TIE stress-test intervals instead resample shared scenarios. Intervals are 95% percentile intervals, and policy differences are paired before the corresponding aggregation. Appendix A.6 maps each result to its recorded artifact.

6 Results

6.1 Covariance gains depend on task and prototype geometry

Table 1 compares covariance-aware gains across representation and prototype types. The visual-prototype rows report the best covariance-aware candidate in the clean track, attained by within-class covariance operators, from $+3.2$ points for CIFAR-100 with DINOv2 to $+25.1$ points with MAE. WebVision reports observed-covariance whitening, with a $+0.72$-point paired bootstrap gain. True text zero-shot CIFAR-100 selects the isotropic endpoint and changes accuracy by 0.00 points across five prompt sets. These controls establish the task-dependent value of covariance candidates.

6.2 Guarded selection improves robust performance without group labels

The central experiment evaluates deployment policies on all 156 scenarios. Table 2 separates average from robust outcomes. Naive selection deploys the candidate with the highest validation-average accuracy from the full non-oracle pool. It improves average accuracy by $+2.5$ points but has a robust delta of -0.6 points, with 30 harmful selections. Guarded-Repair selects within the family-guarded pool of Algorithm 1. Its average gain is $+1.3$ points, while robust gain is $+3.7$ points with

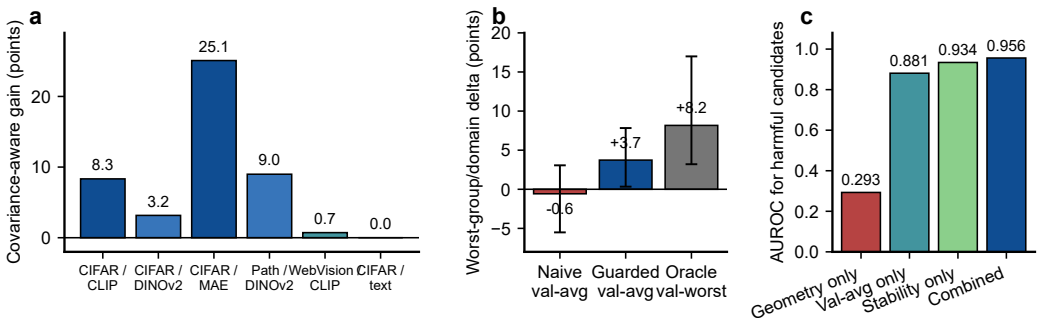

Figure 1: Three complementary audit results. (a) Covariance-aware gains from Table 1: visual-prototype best candidates, WebVision observed whitening, and the text zero-shot cosine endpoint. (b) Across 156 scenarios, family-guarded selection improves robust accuracy; error bars are 95% bootstrap intervals over dataset means. (c) The separate combined risk diagnostic reaches AUROC 0.956 on observed-covariance candidates.

Table 2: Main policy comparison over 156 scenarios. Deltas are percentage points relative to cosine; CI refers to robust delta. The guarded row uses Algorithm 1; the oracle uses group labels for validation selection.

Policy	Avg. Δ	Robust Δ	95% CI	Improve / Worse
Naive val-avg	$+2.5$	-0.6	$[-5.5, +3.1]$	$79/30$
Guarded val-avg	$+1.3$	$\mathbf{+3.7}$	$[+0.3, +7.8]$	$\mathbf{85/9}$
Oracle val-worst	$+0.4$	$+8.2$	$[+3.2, +17.0]$	$114/7$

a 95% interval of $[+0.3, +7.8]$, and harmful selections fall to 9. The paired guarded-minus-naive robust difference is $+4.3$ points with interval $[+0.0, +12.9]$ and is positive in 21 paired scenario comparisons.

The oracle selects by validation worst-group/domain accuracy. Its $+8.2$-point robust gain provides a group-informed comparison. Family-guarded selection trades some average accuracy for improved robust outcomes: its paired average-accuracy difference from naive selection is -1.3 points. The result supports an explicit deployment trade-off between average gain and harmful selections.

6.3 Cross-dataset behavior and the value of explicit nuisance repair

Table 3 resolves the aggregate result into shift families. Waterbirds is the sharpest test of group-free selection: the guarded policy improves worst-group accuracy by $+0.120$ for CLIP and $+0.111$ for SigLIP, while naive selection decreases it by -0.135 and -0.065, respectively. These gains survive class-only fit-subset construction. CelebA gains are positive for both backbones, with an all-backbone robust delta of $+0.042$. PACS improves worst-domain accuracy by $+0.024$ and confirms that the policy extends beyond a binary attribute benchmark.

The synthetic shifts identify where the observed benefits stop. Border corruption yields $+0.022$, whereas tint aggregates to -0.016 across the evaluated strengths and seeds, with a wide interval. Explicit nuisance candidates and validation-average selection therefore do not resolve every shift in the tested suite.

6.4 Risk diagnostics explain why geometry alone is insufficient

We next evaluate the candidate-risk signals of Section 4.4. The observed-covariance subset contains 666 deployable candidates, of which 521 are harmful under the robust metric. In Table 4, VAI geometry alone has AUROC 0.293 and a negative Spearman association with negative robust delta. Validation-average preservation reaches 0.881, and prompt stability reaches 0.934. Combined risk

Table 3: Family-guarded policy by dataset and backbone. Robust metric is worst-group except for PACS, where it is worst-domain. Deltas are percentage points; single-backbone CIs resample seed-averaged clusters, and *both* CIs resample backbone means.

Dataset	Backbone	Robust Δ	95% CI
Waterbirds	CLIP	$+12.0$	$[+1.4, +26.4]$
Waterbirds	SigLIP	$+11.1$	$[+0.0, +29.9]$
CelebA	both	$+4.2$	$[+3.3, +5.0]$
PACS	both	$+2.4$	$[+2.0, +2.7]$
CIFAR border	CLIP	$+2.2$	$[+0.2, +5.8]$
CIFAR tint	CLIP	-1.6	$[-7.3, +2.1]$

Table 4: Harmful-candidate prediction on 666 observed-covariance candidates. Harmful means negative robust delta. Spearman correlates risk with negative robust delta; bold marks the best value in each column.

Predictor	AUROC	AUPRC	Spearman
Geometry only	0.293	0.678	-0.492
Validation average only	0.881	0.973	$\mathbf{+0.915}$
Prompt stability only	0.934	**0.994**	$+0.764$
Validation + geometry	0.784	0.797	$+0.532$
Combined risk	**0.956**	0.993	$+0.762$

has the highest reported AUROC, 0.956, with AUPRC 0.993. On the broader deployable non-noop pool, the fixed accept/reject decision has 0.761 accuracy and recalls 0.854 of harmful candidates.

These results position VAI as a conditional diagnostic. Its spectrum-level interpretation complements empirical validation and prompt evidence, while the combined score is strongest in AUROC on this candidate family. Prompt stability has the highest AUPRC among the reported rows. This analysis measures harmful-candidate detection separately from the main policy's deployment gains.

6.5 External TIE candidates strengthen the audited pool

The external translation-style candidate moves image embeddings along a spurious text direction. We test this mechanism through a feature-cache adaptation on shared Waterbirds and CelebA scenarios. This experiment uses the fixed signal gate of Section 4.4 on the full base candidate pool, TIE-style candidates, or their union, with cosine always available. It evaluates pool complementarity under a common gate, separately from the main family guard.

Table 5 and Figure 2 show two benefits. Under the fixed signal gate, TIE-only candidates yield $+5.4$ robust points and the base pool yields $+3.2$. Their union reaches $+7.3$ points with 95% interval $[+4.9, +9.7]$, a paired gain of $+4.1$ points over the base pool under the same gate, with interval $[+2.5, +6.1]$. Within the enlarged pool, the fixed gate selects 9 harmful candidates, compared with 18 for validation-average selection. External candidates improve the pool's robust gain, while auditing reduces harmful deployments from that pool.

6.6 Negative controls and selection boundaries

A ZERO-style no-backpropagation TTA control uses 16 image augmentations with mean logits or majority voting. Across five prompt sets on Waterbirds, center evaluation has average accuracy 0.668 and worst-group accuracy 0.298. Mean-logit aggregation changes these to 0.673 and 0.295, while voting gives 0.673 and 0.297. This control improves average accuracy slightly without improving worst-group robustness, reinforcing the need to evaluate both outcomes.

We also evaluate held-out threshold transfer for the direct signal-gate diagnostic. Thresholds are chosen using robust test outcomes on non-held-out benchmark families and then evaluated on the held-out family. The selected VAI threshold is 0.10 for CelebA, CIFAR border, CIFAR tint, and PACS, and 0.05 for Waterbirds. Held-out robust deltas are $+0.042$, $+0.022$, -0.016, and $+0.024$, respectively; Waterbirds yields $+0.004$. Holding out the backbone selects thresholds between 0.075

Table 5: TIE-style stress test over 120 shared Waterbirds and CelebA scenarios. *Base* denotes our full non-oracle candidate pool; *fixed gate* uses Section 4.4. Robust deltas are percentage points, with scenario-bootstrap CIs.

Policy	Robust Δ	95% CI	Harmful selected
Base only, fixed gate	+3.2	[+1.1, +5.3]	7
TIE only, fixed gate	+5.4	[+3.7, +7.2]	7
Base + TIE, fixed gate	**+7.3**	[+4.9, +9.7]	9
Base + TIE, val-avg	+7.0	[+3.8, +10.1]	18

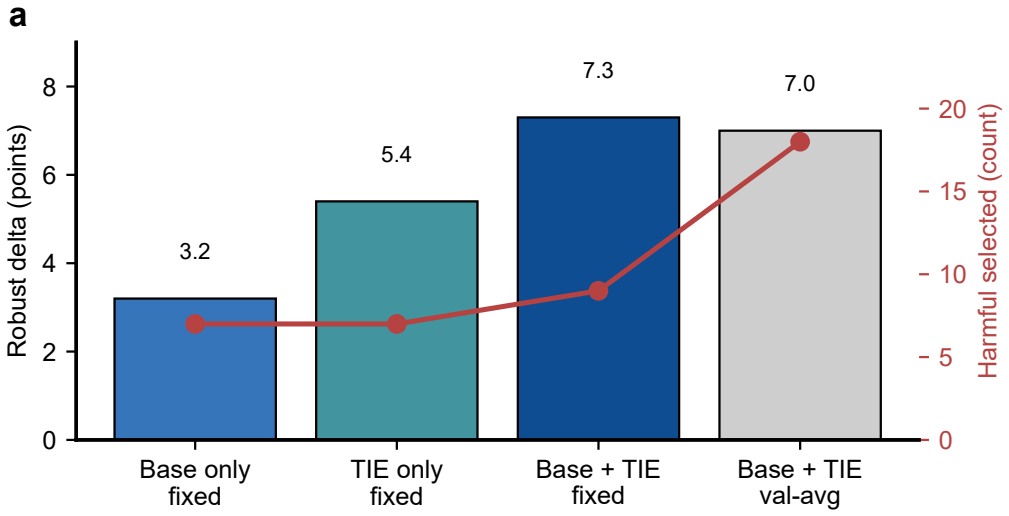

Figure 2: External-pool complementarity and selection risk on 120 shared scenarios. Bars show robust delta; the red line counts harmful selections. The combined pool has the highest robust point estimate under the fixed signal gate, which reduces harmful selections relative to validation-average selection from the same pool.

and 0.10 and yields +0.050 for CLIP and +0.016 for SigLIP. These are cross-benchmark meta-validation diagnostics; they do not set the main family-guarded policy.

7 Discussion

The experiments establish candidate selection as a substantive part of VLM repair. Covariance candidates help visual-prototype tasks, the text zero-shot control selects the cosine endpoint, and explicit nuisance repairs improve several shifted settings. A deployment decision must account for this variation across candidate families and tasks.

The main family guard makes this principle operational using class-labeled validation accuracy and text or pseudo nuisance directions, without protected-attribute labels for fitting or selection. The separate signal audit adds prompt stability and variance geometry to diagnose candidate risk. The cross-backbone results and TIE stress test show how selection and diagnosis work alongside existing repair mechanisms.

Variance alignment also characterizes low-light failure cases. Direct text-axis VAI is negative where raw Mahalanobis accuracy (Lee et al., 2018) falls from 88.2% to 9.1% on VOC-object, from 76.6% to 9.7% on ExDark, and from 72.7% to 33.3% on BDD-night. VAI alone nevertheless has AUROC 0.293 on observed-covariance candidates. Its role is to describe variance geometry within a broader empirical audit.

The tested scope includes Waterbirds and CelebA group shifts, PACS domain variation, and synthetic border and tint shifts. CIFAR tint remains weak under the family guard, and its interval spans zero. TIE is evaluated through a feature-cache adaptation. These observed boundaries motivate richer nuisance candidates and online monitoring, extending the same audit-and-selection formulation.

An important consequence is that the audit produces an interpretable deployment trace rather than a single repaired representation. For each scenario, the recorded trace identifies the candidate family, prompt set, fit seed, covariance settings, validation-average change, prompt-pass rate, and applicable signal checks. This makes it possible to attribute a robust gain to candidate generation or to deployment selection. In the main suite, the family guard changes the policy outcome while retaining cosine in the eligible pool; in the TIE stress test, the signal checks are held fixed while the candidate pool changes. The comparison therefore isolates pool complementarity from the gate itself. This separation also explains why VAI is reported as a diagnostic ranking and not as a replacement objective for validation selection.

The results give each repair mechanism a concrete operational role. Covariance-aware candidates are useful to test on visual-prototype tasks, while the text zero-shot control prevents whitening from being treated as universally beneficial. The group-shift tables show where explicit nuisance-direction families help, and the tint and low-light cases identify settings in which a candidate can lose robust accuracy despite improving an average or raw score. The framework consequently turns heterogeneous repair mechanisms into comparable deployment choices with a recorded evidence trail, while keeping the robust cost of a poor choice visible.

This perspective changes how a repair result should be reported. A single test accuracy for the selected candidate is insufficient because it does not reveal which candidates were considered, which validation signal selected the winner, or whether a candidate was rejected for instability. Our protocol records these decisions at the scenario level and then aggregates them with the same units used for uncertainty estimation. The resulting report separates three questions that are often conflated: whether a repair family can produce a useful candidate, whether an audit can reject candidates that are risky on the tested shift, and whether an expanded pool supplies complementary choices. The main suite addresses the first two questions under a common family guard; the TIE experiment addresses the third under a fixed gate. This organization makes the empirical comparison legible even when the best candidate differs across datasets, backbones, prompt sets, and shift types.

The framework is therefore most useful as an evaluation layer around existing frozen-encoder methods. It does not require replacing a repair mechanism or retraining the VLM. Instead, it makes the deployment choice explicit, preserves a cosine fallback, and exposes cases in which average validation improvement is a poor proxy for robust behavior. The negative controls and failure-boundary analyses are part of this operational statement: they show that a repair can be computationally valid and still be unsuitable for a particular shift. In this sense, the contribution is a reproducible decision protocol for comparing repairs under the same evidence rules, rather than a claim that one transformation dominates all others.

8 Conclusion

We formulated frozen VLM repair as an audit and candidate-selection problem. Across 156 scenarios, the group-free family guard changes a -0.6-point robust delta from naive selection to $+3.7$ points and reduces harmful selections from 30 to 9. A separate combined risk diagnostic reaches 0.956 AUROC on observed-covariance candidates. On shared Waterbirds and CelebA scenarios, the TIE-augmented pool reaches $+7.3$ robust points under the fixed signal gate, improving over the base pool under that gate by $+4.1$ points. Repair methods supply candidates; auditing and selection determine their deployment value. This division of labor enables robust selection to improve alongside the repair methods it evaluates.

Reproducibility Statement

Appendix A.6 identifies the recorded artifacts behind the reported results, and Appendix A.7 describes the per-scenario decision trace. The evaluation records dataset, backbone, prompt set, fit

seed, covariance rank, shrinkage coefficient, and policy. Group labels are excluded from deployable fitting and selection; benchmark construction and diagnostic uses are stated in the protocol. The implementation uses frozen encoders and deterministic inference for each recorded seed. The submission source package includes the manuscript and vector figures; it does not include experiment scripts, feature caches, or model weights.

ETHICS STATEMENT

This work uses public image benchmarks and studies model-selection behavior under group and domain shifts. It does not collect human-subject data or infer sensitive attributes for individual deployment. Protected or nuisance labels are used only for benchmark construction, evaluation, and explicitly identified oracle or diagnostic comparisons; deployable fitting and selection exclude those labels. The reported risk is therefore a property of the evaluated benchmark protocols, not a guarantee of fairness in an untested application.

AI USE STATEMENT

Language models assisted with drafting and editing. All experimental values, algorithmic definitions, citations, and claims were checked against the project artifacts and verified bibliographic metadata by the authors.

REFERENCES

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form, 2023. URL https://arxiv.org/abs/2306.03819.

Konstantinos M. Dafnis and Dimitris N. Metaxas. Test-time spectrum-aware latent steering for zero-shot generalization in vision-language models, 2025. URL https://arxiv.org/abs/2511.09809.

Sepehr Dehdashtian, Lan Wang, and Vishnu Naresh Boddeti. Fairerclip: Debiasing clip's zero-shot predictions using functions in rkhss, 2024. URL https://arxiv.org/abs/2403.15593.

Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models, 2024. URL https://arxiv.org/abs/2405.18330.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Walter Gerych, Haoran Zhang, Kimia Hamidieh, Eileen Pan, Maanas Sharma, Thomas Hartvigsen, and Marzyeh Ghassemi. Bendvlm: Test-time debiasing of vision-language embeddings, 2024. URL https://arxiv.org/abs/2411.04420.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2007.01434.

Jameel Hassan, Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization, 2024. URL https://arxiv.org/abs/2311.01459.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022. doi: 10.1109/CVPR52688.2022.01553.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5637–5664. PMLR, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks, 2018. URL https://arxiv.org/abs/1807.03888.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization, 2017a. URL https://arxiv.org/abs/1710.03077.

Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017b. URL https://arxiv.org/abs/1708.02862.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015. doi: 10.1109/ICCV.2015.425.

Maxime Oquab, Timothee Darcet, Theo Moutakanni, Huy Vo, Marc Szafraniec, et al. Dinov2: Learning robust visual features without supervision, 2023. URL https://arxiv.org/abs/2304.07193.

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, 2021. URL https://arxiv.org/abs/2103.00020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020. URL https://arxiv.org/abs/1911.08731.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems 35*, pp. 14274–14289, 2022. doi: 10.52202/068431-1038. URL https://arxiv.org/abs/2209.07511.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2006.10726.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2 – a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 2023. doi: 10.1038/s41597-022-01721-8. URL https://arxiv.org/abs/2110.14795.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D. Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion, 2024. URL https://arxiv.org/abs/2403.14119.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. doi: 10.1109/ICCV51070.2023.01100.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *Computer Vision – ECCV 2022*, pp. 493–510. Springer Nature, 2022. doi: 10.1007/978-3-031-19833-5_29. URL https://arxiv.org/abs/2207.09519.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16795–16804. IEEE, 2022a. doi: 10.1109/CVPR52688.2022.01631. URL https://arxiv.org/abs/2203.05557.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models, 2022b. URL https://arxiv.org/abs/2109.01134.

Beier Zhu, Jiequan Cui, and Hanwang Zhang. Robust fine-tuning of zero-shot models via variance reduction, 2024a. URL https://arxiv.org/abs/2411.06966.

Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, and Hanwang Zhang. Enhancing zero-shot vision models by label-free prompt distribution learning and bias correcting, 2024b. URL https://arxiv.org/abs/2410.19294.

A Additional Protocol Details

A.1 Candidate pool

The full deployable audit pool includes cosine, observed-covariance whitening and Mahalanobis scoring, top-variance filtering, text and pseudo nuisance-direction filtering, and domain-direction filtering for PACS. The main family guard admits only cosine and the explicit text/pseudo nuisance or domain families. Oracle candidates use true nuisance labels for a group-informed comparison. Candidate complexity is ordered as cosine, observed-covariance operators, text nuisance filters, pseudo nuisance filters, and oracle operators, after validation-average ties. The external-pool experiment applies its fixed signal gate to the full base pool and TIE candidates, rather than applying the main family restriction.

A.2 Prompt and split controls

Waterbirds uses five fixed class-prompt sets and three fit-subset seeds (123, 456, and 789) across the two backbone tracks. The class-only fit subset contains 75% of the training data and is stratified by class without using place labels. CelebA uses three correlation levels, three fit seeds, and five class-prompt sets. PACS uses three prompt sets and deterministic class-domain-stratified 60/20/20 splits for seeds 123, 456, and 789. The synthetic CIFAR suite uses border strengths 2, 6, and 10 pixels and tint alpha values 0.25, 0.48, and 0.70 with seeds 123, 456, and 789.

A.3 External baseline protocol

FairerCLIP is reported from its official local implementation on prepared feature caches, with labeled and no-label variants on Waterbirds and CelebA. LEACE is evaluated as a closed-form feature-cache baseline. The TIE-style translation candidate is evaluated on the same cached CLIP and SigLIP features. ZERO is evaluated as a no-backpropagation multi-view control because its official loaders do not include Waterbirds or CelebA. These protocol distinctions are shown in the baseline status artifact and prevent a feature-cache stress test from being presented as an official end-to-end rerun.

A.4 Compute and execution

Canonical artifacts were generated on a server using Python 3.11.15, PyTorch 2.11.0+cu128, and CUDA 12.8 on two NVIDIA GeForce RTX 5090 D v2 GPUs (24,455 MiB each). Feature extraction and candidate scoring use one GPU per process; policy aggregation and bootstrap summaries operate on recorded feature caches. The experiment repository's result index records the script, dataset, backbone, prompt set, fit split, seed, covariance rank, shrinkage coefficient, and policy for each row.

The submission package documents the protocols and artifact identities but does not include the experiment repository, feature caches, or model weights.

A.5 Observed-covariance failure boundary

For low-light recognition, the image features are paired with true text prompts for VOC-object, ExDark, and BDD-night. The direct semantic and nuisance axes are projected onto the observed covariance eigenbasis. Raw inverse covariance corresponds to the zero-shrinkage endpoint. The validation-selected shrinkage endpoint returns to the isotropic cosine geometry, which is why the failure is reported as a boundary of raw Mahalanobis rather than as a failure of the guarded policy.

A.6 Result provenance and policy separation

All paths below are relative to the experiment repository's `outputs/` directory. Tables 2 and 3, and Figure 1b, use `unified_policy_selection_v6_groupfree_waterbirds_with_pacs/policy_aggregate_ci.csv`. The reported guarded row is `deploy_nuisance_guarded_valavg`, whose family restriction is defined in Algorithm 1.

Table 4 uses `paper_ready_artifacts_v2/observed_cov_family_metrics_table.csv`, derived from `candidate_repair_risk_validation_v2/`. Its fixed accept/reject and held-out threshold checks concern the signal-gate diagnostic, not the main family-guarded policy. Table 5 uses `paper_ready_artifacts_v2/external_tie_stress_table.csv`; its paired comparison uses `external_tie_pairwise_table.csv` in the same directory. These summarize `tie_external_candidate_gate_audit_v2/`, with one selected candidate per shared scenario and scenario-bootstrap intervals.

Table 1 uses `generic_cov_pilot_multiseed/summary.md`; WebVision uses `webvision_cov_pilot/bootstrap_ci.md`. Text controls and low-light diagnostics use `prompt_sensitivity/summary.md` and `lowlight_text_axis_diagnostics/summary.md`. ZERO-style controls use `baseline_comparison_v3/zero_style_waterbirds.csv`.

A.7 Aggregation and decision trace

For every scenario, the recorded candidate row contains the dataset and backbone identifiers, prompt set, fit or split seed, candidate family, candidate parameter, validation-average score, robust test outcomes, and the policy that selected it. The main aggregate first averages repeated seeds within the cluster defined in Section 5; combined-backbone rows then resample backbone means, and the overall row resamples equally weighted dataset means. Harmful selection is counted only after the selected candidate is compared with the cosine baseline on the scenario's robust metric. The fixed gate records each applicable check separately, so a rejected candidate can be distinguished from a candidate that was never eligible for a given geometry test. These records support the policy separation used in the main text without exposing group labels to deployable fitting or selection.