RELATIONAL STRUCTURE RECOVERY FOR STRICT SCORE-ONLY VLM ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deployment interfaces for vision-language models may expose only a fixed tensor of prompt-level class scores, withholding images, embeddings, gradients, parameter updates, labels, and additional queries. Existing score-only methods mainly aggregate or reweight these scores. We ask instead whether the tensor contains a relational representation that supports target-conditioned adaptation. Prompt-class scores are repeated text-induced measurements: on non-degenerate rows, normalized class contrasts induce a shared geometry and form a maximal invariant to sample-wise offsets and positive scales, while prompt-dependent deviations provide complementary information after the shared response is accounted for. Prompt-Class Residual Interaction (PCRI) instantiates this shared–conditional decomposition by constructing a Score-space Core, extracting Core-conditioned prompt innovation, and using the joint representation only to correct the frozen semantic score field through target prototypes. With one configuration selected on a disjoint CIFAR-100 development setting and frozen across all evaluation domains, PCRI improves matched CARPRT by +2.378 percentage points on average across 11 datasets and three VLM backbones, with gains in 32/33 settings under identical cached scores. Matched controls attribute most Core-only gain to class-contrast standardization and isolate a further +0.306-point benefit from Core conditioning; support stress tests show that limited or unrepresentative target support can make the correction unreliable. PCRI uses no hidden representations or additional VLM calls after score extraction.

1 INTRODUCTION

Pre-trained vision-language models (VLMs) enable zero-shot recognition, but their predictions remain prompt-sensitive (Radford et al., 2021), motivating score-based prompt weighting such as WPE and CARPRT (Allingham et al., 2023; Dong et al., 2026). Most test-time adaptation methods instead require gradients, internal features, repeated evaluations, or learnable representations (Shu et al., 2022; Karmanov et al., 2024; Zanella & Ben Ayed, 2024; Farina et al., 2024; Zhou et al., 2025; Wu et al., 2026). We study the stricter fixed-pool interface: after prompt-level class scores are cached, adaptation cannot access images or embeddings, update parameters, use gradients or labels, or query the VLM again.

This setting poses a representation question beyond score aggregation: *what relational information is recoverable from prompt-class scores alone, and how can it support adaptation without replacing pretrained class semantics?* Our observation is that prompts provide repeated *text-induced measurements*. Normalized class contrasts induce a shared relational geometry invariant to sample-wise common offsets and positive scales. Prompt-dependent variation is complementary only after removing the part predictable from this shared response. The tensor thus admits a shared–conditional decomposition: common class-contrast geometry and prompt-conditioned innovation beyond it.

PCRI instantiates this principle with a Score-space Core, Core-conditioned prompt innovation, and a scale-matched correction anchored to frozen semantic scores. The novelty lies in this access-matched decomposition and its boundary; normalization, ridge, prototypes, and fusion are implementation operators rather than separate claims. Figure 1 provides post-hoc geometric diagnostics, with broader quantitative evidence in Appendix D.3 and Figure 3.

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.

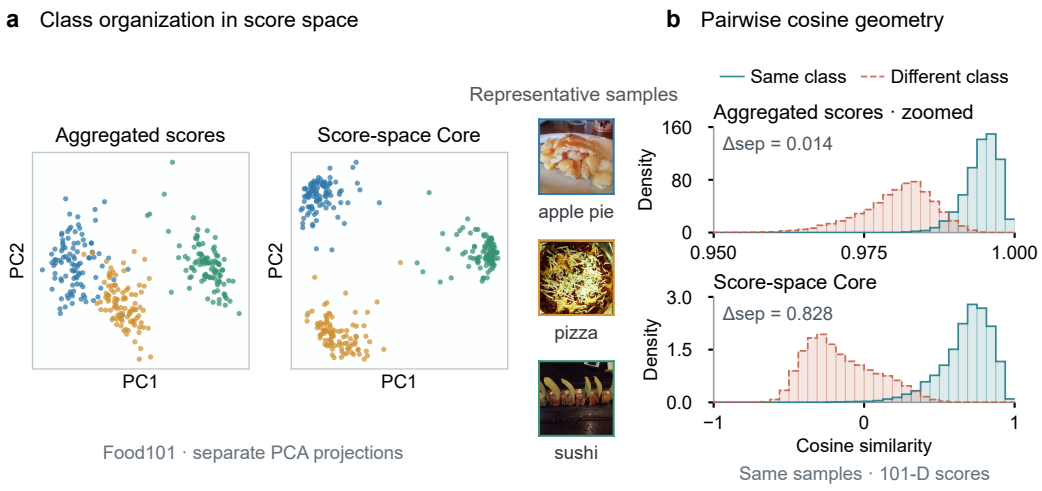


Figure 1: Class geometry in cached score space. (a) Separate deterministic 2D PCA projections of the same 300 Food101 images compare within-representation organization in B and U, not motion in a shared coordinate system; photographs identify the three displayed classes. (b) Empirical density histograms of same-class and different-class cosine similarities are computed in all 101 score coordinates. The raw-score axis is zoomed to $[0.95, 1]$, whereas the Core axis spans $[-1, 1]$. Δ_{sep} is the unbinned mean cosine gap, not an accuracy gain. Labels and photographs are used only for post-hoc illustration; Appendix D.3 gives sampling, pair counts, and binning.

Across 11 datasets and three VLM backbones, PCRI improves matched CARPRT by +2.378 percentage points on average and improves 32 of 33 settings when evaluated on identical frozen score tensors. Five-fold query-excluded cross-fitting retains 97.6% of this gain, while mechanism controls, difficulty stratification, and target-support stress tests probe the explanation and delineate the operating regime.

Contributions. (1) We formulate strict score-only adaptation as relational recovery from repeated prompt-class measurements and characterize normalized class contrasts as a maximal invariant under sample-wise positive-affine nuisance. (2) PCRI separates shared class-contrast geometry from conditional prompt innovation and anchors target correction to frozen semantics. (3) We formalize the score-only observability boundary: cached scores alone cannot guarantee distribution-free sample-wise no-harm, and classes absent from base pseudo-label support are not structurally identifiable. Cross-fitting and support stress tests delimit this regime across 33 matched settings.

2 RELATED WORK

2.1 PROMPT AND SCORE-BASED VLM ADAPTATION

Prompt learning improves VLM transfer through learnable context, prompt distributions, multimodal or visual prompts, and generated descriptions (Zhou et al., 2022; Lu et al., 2022; Jia et al., 2022; Khattak et al., 2023; Pratt et al., 2023). Training-free methods exploit external support or prediction statistics (Udandarao et al., 2023). WPE estimates prompt quality without labeled validation data; CARPRT learns class-specific weights from unlabeled scores (Allingham et al., 2023; Dong et al., 2026). These methods adapt *prompt aggregation*; PCRI instead treats prompt scores as repeated measurements, quotients a specified nuisance group from their shared response, and tests whether pre-aggregation variation remains useful conditional on that response.

2.2 RESTRICTED TEST-TIME ADAPTATION

Classical test-time adaptation modifies model statistics, parameters, or classifiers from unlabeled deployment data (Liang et al., 2020; Wang et al., 2021; Niu et al., 2022; 2023). VLM-specific methods similarly rely on differentiable prompt tuning, augmented evaluations, visual-feature caches, mode seeking, prediction marginalization, or adapted class representations (Shu et al., 2022; Feng et al.,

2023; Samadh et al., 2023; Karmanov et al., 2024; Zanella & Ben Ayed, 2024; Farina et al., 2024; Zhou et al., 2025; Wu et al., 2026). Query-based black-box prompt tuning such as B^{2}TPT removes gradient access but still optimizes prompts through repeated inference-API queries and derivative-free search (Meng et al., 2025); PCRI instead moves adaptation to a fixed observable space: after the prompt-class score tensor is materialized, no further VLM access is required.

Prototype and transductive methods show that class representatives can support relational prediction when a meaningful representation is available (Snell et al., 2017; Ziko et al., 2020; Boudiaf et al., 2020). PCRI does not claim standardization or prototypes as novel; Section 5 directly tests whether raw-score prototypes already explain the gain and whether pre-aggregation information remains conditionally useful.

Because PCRI estimates its state from the unlabeled pool, limited or unrepresentative support is an explicit operating boundary (Section 5; Appendix F), consistent with realistic TTA concerns (Zanella et al., 2025).

3 PROBLEM SETUP

Let the fixed unlabeled target pool contain N images and C candidate classes, with a common bank of P textual prompts. A single extraction phase materializes the prompt-class score tensor $S \in \mathbb{R}^{N\times P\times C}$, with $\tau = \exp(\gamma)$. A fixed score-only prompt aggregator $\mathcal{A}$ maps this tensor to base class scores:

$$S[n,p,c] = \tau\, \bar{t}_{p,c}^{\mathsf{T}} \bar{z}_n, \qquad B = \mathcal{A}(S) \in \mathbb{R}^{N\times C}. \tag{1}$$

The base prediction is $\hat{y}_n^{\text{base}} = \arg\max_c B[n,c]$. The normalized image/text representations $\bar{z}_n, \bar{t}_{p,c}$ are internal to extraction and unavailable to PCRI. This measurement interpretation assumes neither an orthogonal text basis nor hidden-feature reconstruction. We use CARPRT as $\mathcal{A}$, with MPE and WPE in the aggregator-generality analysis.

Strict score-only transduction. After S is materialized, PCRI may use only cached scores and sample/prompt/class order, with no embeddings, activations, parameters, gradients, target labels, or further VLM queries. The map $\mathcal{F}_{\text{PCRI}} : S \mapsto \hat{y}$ estimates one state jointly from the complete unlabeled pool; we therefore claim only the fixed-pool target-transductive regime, not online, continual, or single-sample adaptation.

Method and hyperparameters are frozen using CIFAR-100/CLIP ViT-B/16 development labels only. Evaluation labels are read after predictions are fixed. Matched CARPRT and PCRI share the frozen tensor and sample/prompt/class order (Appendix B).

4 RELATIONAL STRUCTURE RECOVERY

Overview. PCRI recovers a nuisance-invariant Core, retains prompt variation less predictable from it, and uses the joint representation only to correct frozen base scores.

4.1 SHARED CLASS-CONTRAST GEOMETRY

Raw scores B rank classes rather than define cross-sample distances; common offsets and positive rescalings can obscure relative responses. We interpret each aggregated class response as a linear measurement of the hidden visual representation. For fixed weights, let $M = [m_1, \ldots, m_C]$. The effective class measurements and image response satisfy

$$m_c = \sum_{p=1}^{P} w_{p,c} \bar{t}_{p,c}, \qquad b_n = \tau M^{\mathsf{T}} \bar{z}_n. \tag{2}$$

and let $P_C = I - \frac{1}{C}\mathbf{1}\mathbf{1}^{\mathsf{T}}$ center responses across classes.

For a non-degenerate score row, let $\mathcal{X} = \{b \in \mathbb{R}^C : P_C b \neq 0\}$ and define $q(b) = P_C b / \|P_C b\|_2$.

Proposition 1 (Induced geometry and maximal invariance). For fixed M: (i) centered score responses induce the positive-semidefinite bilinear similarity

$$(P_C b_i)^{\mathsf{T}}(P_C b_j) = \tau^2 \bar{z}_i^{\mathsf{T}} K_M \bar{z}_j, \qquad K_M := M P_C M^{\mathsf{T}} \succeq 0. \tag{3}$$

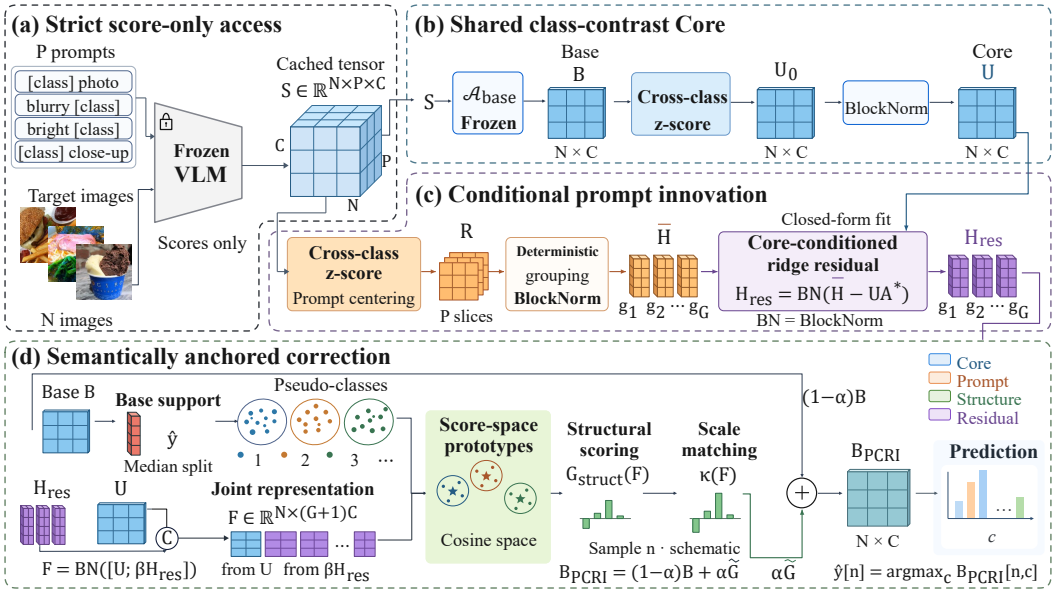


Figure 2: Shared–conditional relational recovery under strict score-only access. (a) PCRI receives only cached prompt–class scores. (b) Normalized class contrasts form the shared Core U. (c) Core conditioning retains prompt innovation H_{res}. (d) The joint representation supplies target relations, while frozen scores supply pseudo-class semantics and anchor the scale-matched correction.

and (ii) q is a maximal invariant on $\mathcal{X}$ for the nuisance group $\mathcal{G} = \{b \mapsto ab + \delta\mathbf{1} : a > 0, \delta \in \mathbb{R}\}$:

$$q(b') = q(b) \iff \exists\, a > 0, \delta \in \mathbb{R} :\ b' = ab + \delta\mathbf{1}, \qquad U_0^\circ(b) = \sqrt{C}\, q(b). \tag{4}$$

Thus U_0° retains exactly the information distinguishable modulo the stated group. This representation property neither models every real VLM shift nor implies accuracy improvement. Appendix A.2 gives the proof, safeguards, and a conditional bounded-distortion result.

The implementation uses the safeguarded U_0 in Eq. (A6), which equals U_0° whenever its variance floor is inactive. PCRI then sets $U = \mathrm{BlockNorm}(U_0)$ as the target-normalized operational Core; we do not claim that this additional dataset-level map is itself maximal. Redundant text directions can limit observable distinctions, so PCRI claims neither hidden-feature recovery nor guaranteed improvement.

4.2 Prompt-conditioned innovation beyond the Core

Aggregation from $S \in \mathbb{R}^{N\times P\times C}$ to $B \in \mathbb{R}^{N\times C}$ suppresses prompt-conditioned variation. Let $Z[n,p,:] = z_{\text{class}}(S[n,p,:])$ apply Eq. (4) to each sample–prompt response, then remove the sample-wise prompt mean,

$$R[n,p,c] = Z[n,p,c] - \frac{1}{P}\sum_{q=1}^{P} Z[n,q,c]. \tag{5}$$

Signed residuals are assigned by a deterministic, non-semantic prompt partition and group-averaged into $H \in \mathbb{R}^{N\times GC}$ (Appendix A.3). This partition provides reproducible compression without encoding prompt semantics; it is neither semantic clustering nor a Johnson–Lindenstrauss projection. Fixed random partitions perform similarly (Appendix E.4). We then normalize this view:

$$\bar{H} = \mathrm{BlockNorm}(H). \tag{6}$$

Because $\bar{H}$ still duplicates variation predictable from U, PCRI removes its ridge-predictable component:

$$A^* = \arg\min_{A\in\mathbb{R}^{C\times GC}} \|\bar{H} - UA\|_F^2 + \lambda\|A\|_F^2 = (U^\top U + \lambda I)^{-1} U^\top \bar{H}. \tag{7}$$

Equivalently, define the ridge shrinkage operator

$$\Pi_\lambda(U) = U(U^\top U + \lambda I)^{-1}U^\top, \qquad \bar{H} - UA^* = (I - \Pi_\lambda(U))\bar{H}. \tag{8}$$

The normal equations give $U^\top(\bar{H} - UA^*) = \lambda A^*$; hence for $\lambda > 0$, $\Pi_\lambda(U)$ is shrinkage rather than an orthogonal projector. We form

$$H_{\text{res}} = \text{BlockNorm}(\bar{H} - UA^*), \qquad F = \text{BlockNorm}([U; \beta H_{\text{res}}]) \in \mathbb{R}^{N\times(G+1)C}. \tag{9}$$

We call H_{res} *conditional prompt innovation*: variation less linearly predictable from the Core, without claiming independence or hidden-manifold recovery.

4.3 Semantically anchored target correction

Recovered relations do not determine class semantics: simultaneously permuting unlabeled prototype identities preserves their geometry. Frozen scores B break this class-permutation symmetry, so pseudo-labels and confidence come only from B,

$$\tilde{y}[n] = \arg\max_c B[n, c], \qquad r[n] = \max_c \text{softmax}(B[n])[c]. \tag{10}$$

Within each pseudo-class $\mathcal{I}_c = \{n : \tilde{y}[n] = c\}$, let $r_c^{\text{med}} = \text{median}\{r[n] : n \in \mathcal{I}_c\}$. Support is selected by a class-wise median confidence split:

$$\mathcal{I}_c^+ = \{n \in \mathcal{I}_c : r[n] \geq r_c^{\text{med}}\}. \tag{11}$$

This parameter-free split uses within-class confidence rank and keeps nonempty support for each nonempty pseudo-class. Operator $\mathcal{P}(X; \tilde{y}, r)$ returns sample–prototype cosine scores $G_{\text{struct}}(X)$; empty pseudo-classes remain invalid.

The structural field and base logits have unrelated scales, so PCRI matches their dispersions:

$$\kappa(X) = \frac{\text{Std}(B)}{\max\{\text{Std}([G_{\text{struct}}(X)]_{\text{finite}}), 10^{-12}\}}, \tag{12}$$

Only finite entries contribute. Above the numerical floor, $\text{Std}(\kappa(X)[G_{\text{struct}}(X)]_{\text{finite}}) = \text{Std}(B)$, so α controls correction strength rather than unit mismatch. Core-only applies the operator to U; full PCRI applies it to F. For the full method, $G_{\text{struct}}(F) = \mathcal{P}(F; \tilde{y}, r)$ and $\widetilde{G}(F) = \kappa(F)G_{\text{struct}}(F)$. With $\Delta(F; B) = \widetilde{G}(F) - B$, the corrected scores are

$$B_{\text{PCRI}} = B + \alpha\Delta(F; B) = (1 - \alpha)B + \alpha\widetilde{G}(F). \tag{13}$$

The final prediction is $\hat{y}[n] = \arg\max_c B_{\text{PCRI}}[n, c]$. Class identities, pseudo-labels, and confidence ordering remain anchored to B; the target branch supplies only a dispersion-matched relational correction. A class absent from base pseudo-labels receives no structural prototype, making incomplete coverage an explicit boundary (Appendix A.5). PCRI is training-free and uses no additional VLM queries; Appendices A.4 and G give safeguards and complexity.

Identifiability and operating boundary. The representation cannot recover distinctions absent from S, and its target correction requires representative pseudo-class support. Without assumptions restricting $P(Y \mid S)$, a score-only adapter that changes a base decision cannot guarantee distribution-free sample-wise no-harm: observationally identical scores admit label states favoring either decision (Proposition F.1). This narrow boundary does not rule out beneficial adaptation under additional structure; Sections 5.4 and 6 test where PCRI's target-state estimate becomes unreliable.

5 Experiments

5.1 Setup and main results

We follow the CARPRT benchmark suite: Caltech101, DTD, EuroSAT, FGVC Aircraft, Food101, Flowers102, Oxford-IIIT Pets, Stanford Cars, SUN397, UCF101, and ImageNet, evaluated with CLIP ViT-B/16, CLIP ResNet-50, and DeCLIP ViT-B/32 (Radford et al., 2021; Li et al., 2022), giving 33 dataset-backbone settings. Class names, official evaluation splits, and the 247-template

prompt bank follow the matched CARPRT protocol. The comparison includes MPE (ICML 2021), Majority Vote and WPE (ICML 2023), and CARPRT (ICLR 2026); Table 1 reports the literature-facing comparison, while all paired gains and confidence intervals use a matched CARPRT implementation evaluated from the identical frozen tensors as PCRI.

We freeze $\alpha = 0.50$, $\beta = 0.75$, $G = 4$, and $\lambda = 30$ once on the CIFAR-100/CLIP ViT-B/16 development setting and use them unchanged elsewhere; development-only sweeps are broadly stable around these defaults (Appendix E.5), with no evaluation-dataset tuning. The primary metric is top-1 classification accuracy; aggregate intervals use the dataset-clustered bootstrap described in Appendix B, which preserves the three backbone results of each benchmark as a cluster.

Table 1: Classification accuracy (%) under strict score-only access. PCRI is compared with published score-only baselines across 11 datasets and three VLM backbones. Bold indicates the best result within each backbone.

Method	Caltech101	DTD	EuroSAT	Aircraft	Food101	Flower102	Pets	Cars	SUN397	UCF101	ImageNet	Average
CLIP-ViT-B/16												
MPE (ICML 2021)	92.50	46.88	51.86	21.49	85.34	64.21	79.46	65.21	64.92	67.41	67.59	64.26
Majority Vote (ICML 2023)	93.10	46.75	52.07	22.93	85.60	67.20	81.27	64.93	65.75	68.30	67.98	65.08
WPE (ICML 2023)	93.09	47.04	49.60	23.28	86.14	66.60	82.38	65.93	65.77	68.33	68.28	65.13
CARPRT (ICLR 2026)	**94.16**	48.90	55.56	24.49	**86.31**	71.36	89.13	66.14	66.93	70.41	68.59	67.45
PCRI (Ours)	94.00	**51.48**	**61.51**	**26.28**	86.22	**75.56**	**91.33**	**69.51**	**70.06**	**73.91**	**69.88**	**69.97**
CLIP-ResNet50												
MPE (ICML 2021)	86.41	41.69	30.34	16.05	75.53	56.95	75.98	55.74	59.32	60.06	59.12	56.11
Majority Vote (ICML 2023)	86.79	42.14	28.86	16.29	76.00	60.06	77.29	56.01	60.40	60.87	59.24	56.72
WPE (ICML 2023)	86.65	40.89	30.65	16.11	76.15	58.82	78.43	56.02	59.71	61.53	59.78	56.79
CARPRT (ICLR 2026)	88.46	41.31	36.84	16.88	76.88	65.56	85.69	56.44	61.28	63.66	59.98	59.36
PCRI (Ours)	**89.25**	**43.44**	**38.53**	**18.00**	**76.96**	**70.16**	**88.06**	**59.62**	**64.49**	**66.19**	**61.39**	**61.46**
DeCLIP-ViT-B/32												
MPE (ICML 2021)	94.04	41.63	28.05	7.10	71.71	77.76	76.75	52.22	62.08	57.87	67.01	57.84
Majority Vote (ICML 2023)	94.26	40.29	27.68	7.70	72.34	78.19	77.75	51.87	62.86	58.20	67.24	58.03
WPE (ICML 2023)	94.08	40.97	27.92	7.54	73.15	81.32	80.92	52.21	63.23	58.91	67.97	58.93
CARPRT (ICLR 2026)	94.37	43.31	33.14	**8.76**	**74.15**	82.42	83.28	52.23	64.12	59.57	**68.08**	60.31
PCRI (Ours)	**94.69**	**46.51**	**42.88**	8.52	74.00	**86.72**	**83.78**	**55.14**	**66.55**	**62.99**	67.51	**62.66**

Venue/year labels identify the source of each baseline definition: MPE follows CLIP (Radford et al., ICML 2021); Majority Vote and WPE follow Allingham et al. (ICML 2023); CARPRT is ICLR 2026. Their displayed values are taken from the CARPRT paper. PCRI is the current ICLR 2027 submission and is evaluated under the frozen strict score-only protocol.

Among the methods displayed in Table 1, PCRI attains the highest accuracy in 28 of 33 cells. Against matched CARPRT on identical frozen tensors, it improves 32/33 settings: 62.322% to 64.700%, a mean +2.378 pp (median +2.541; dataset-clustered 95% CI $[+1.554, +3.196]$). All 11 dataset-averaged gains are positive; leave-one-dataset-out means span +2.147 to +2.588 pp (exact two-sided sign test, $p = 0.00098$). The only setting-level decrease is ViT-B/16 on Caltech101, while all three ImageNet gains are positive. Table S4 and Appendix B.6 give complete values and protocols.

5.2 Does the proposed representation explain the gain?

We test three mechanism hypotheses aligned with the design principles. **H1—Response geometry:** the Score-space Core should explain most structural gain beyond raw-score prototypes. **H2—Conditional complementarity:** the same prompt view should help further, with an additional benefit from Core conditioning that weakens when sample/Core correspondence is broken. **H3—Semantic anchoring:** retaining the frozen base score field should outperform structural-only prediction. Section 5.3 then asks whether the anchored correction concentrates on uncertain base predictions.

To isolate conditioning from added dimensionality, we define a matched direct prompt-view control (Direct-H) that uses exactly the same normalized grouped prompt view $\bar{H}$ but removes the ridge-conditioning step:

$$F_{\text{direct}} = \text{BlockNorm}([U; \beta\bar{H}]). \tag{14}$$

All pseudo-labels, prototypes, scale matching, fusion, frozen scores, and hyperparameters are otherwise identical to full PCRI. Thus Direct-H and full PCRI differ only in whether $\bar{H}$ is conditioned on the Core before concatenation.

Table 2: Mechanism-aligned ablation and controls.

Representation construction

Variant / control	**Effect (pp)**
CARPRT	0.000
Raw-score prototype	+0.364
Cross-class standardized U_0	+1.844
Full Core U	+1.847
Core + direct prompt view	+2.072
Full PCRI	**+2.378**

Conditioning / correspondence controls

Comparison	**Effect (pp)**	**95% CI**
Full PCRI − Direct-H	**+0.306**	**[+0.155,+0.454]**
Correct vs shuffled	+1.227	[0.777,1.689]
True-Core vs permuted-Core	+0.228	[0.028,0.409]
Anchored vs structural-only	+1.027	[+0.241,+1.847]

Mechanism progression. A matched 33-setting Core decomposition gives +0.364 pp for raw-score prototype reasoning, +1.844 pp after cross-class standardization (U_0), and +1.847 pp for the full Core U. Thus U_0 improves over the raw control by +1.480 pp (95% CI [+0.946,+2.021]), whereas the subsequent BlockNorm changes the Core-only mean by only +0.003 pp (95% CI [-0.015,+0.027]). The dominant structural gain therefore follows the stated class-contrast quotient rather than prototype reasoning alone. Relative to Core-only, the full prompt branch adds +0.531 pp. Within that increment, Direct-H contributes +0.225 pp and Core conditioning contributes a smaller but independently matched +0.306 pp (95% CI [+0.155,+0.454]). Held-out Core predictability falls from $R^2 = 0.616$ to 0.141, while anchored fusion exceeds structural-only by +1.027 pp. Appendix D gives full controls and protocols.

5.3 Where does the structural correction help?

Figure 3 localizes PCRI's gains to uncertain base predictions: margin gains fall from 7.526 to 0.018 pp and confidence gains from 6.637 to 0.114 pp across hardest-to-easiest quintiles; prompt-group disagreement is weaker but remains positive. Figure 1 illustrates class organization and pairwise cosine geometry in one matched Food101 view, while Appendix D.3 independently quantifies representation suitability across ten datasets.

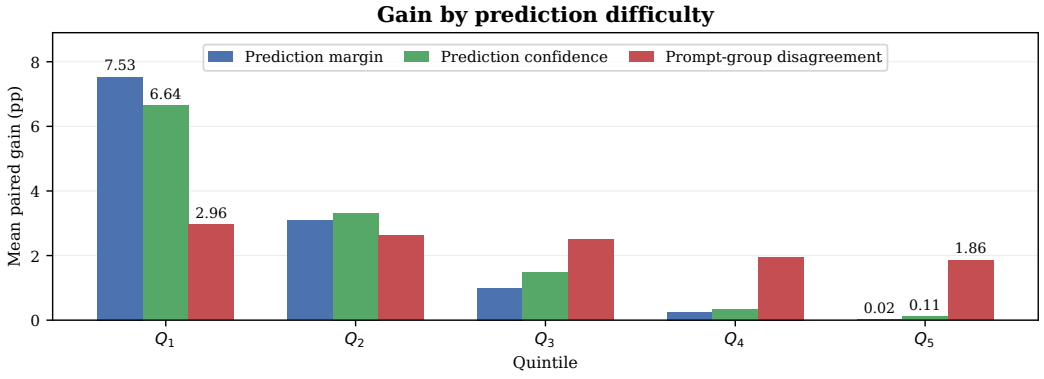


Figure 3: PCRI gains by sample difficulty. Samples are partitioned into quintiles by CARPRT margin, confidence, and prompt-group disagreement; bars report mean PCRI-minus-CARPRT accuracy gain across the 33 matched settings.

5.4 When is target-conditioned geometry reliable?

With the frozen base score field held fixed, Figure 4 varies only the unlabeled target samples available to estimate PCRI's transductive state. Using 10% adaptation support gives -1.239 pp, 25% gives +1.827 pp, and gains change little beyond 50% (+2.482 pp), marking a support requirement rather than a universal sample threshold. Prompt-budget gains remain positive from 16 to 247 prompts (+2.745 to +2.685 pp).

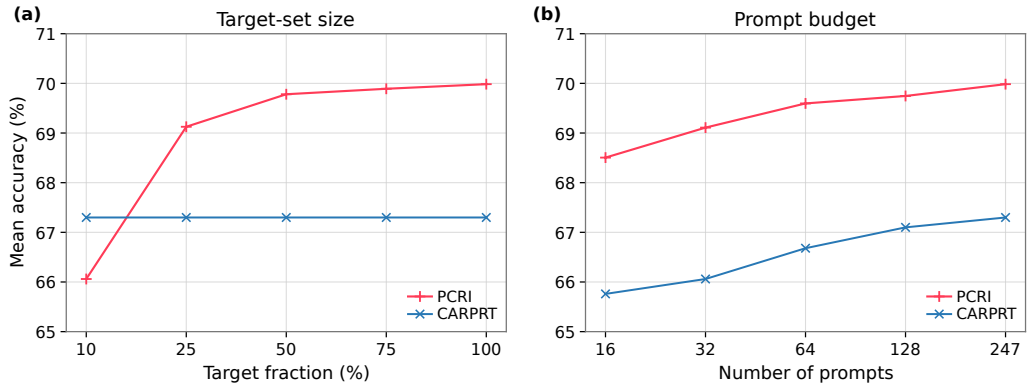


Figure 4: Sensitivity to unlabeled adaptation-support fraction and prompt budget. (a) With frozen base scores, mean gain is negative at 10% support and positive from 25% onward in this tested suite; this is descriptive, not a universal threshold. (b) Gains remain positive across prompt budgets. PCRI accuracies equal matched CARPRT accuracy plus the final paired gain (Appendix E.3); ten datasets, CLIP ViT-B/16, frozen hyperparameters.

5.5 Does the fitted target state transfer to held-out queries?

To test whether a query benefits from contributing to its own fitted target state, we perform five-fold cross-fitting: every query is excluded from CARPRT weighting, centering, ridge conditioning, pseudo-support selection, prototype construction, and scale matching. Across 33 settings, the held-out gain is +2.320 pp (dataset-cluster bootstrap 95% CI [+1.503,+3.120]), retaining 97.6% of the canonical +2.378 pp full-pool gain; 32/33 settings are positive and backbone means are +2.558 pp (ViT-B/16), +2.056 pp (RN50), and +2.348 pp (DeCLIP-ViT-B/32). Full-pool and held-out gains correlate at $r = 0.992$; two additional fold allocations both give +2.277 pp. Table S3 specifies the protocol, and Table S19 indexes every setting row.

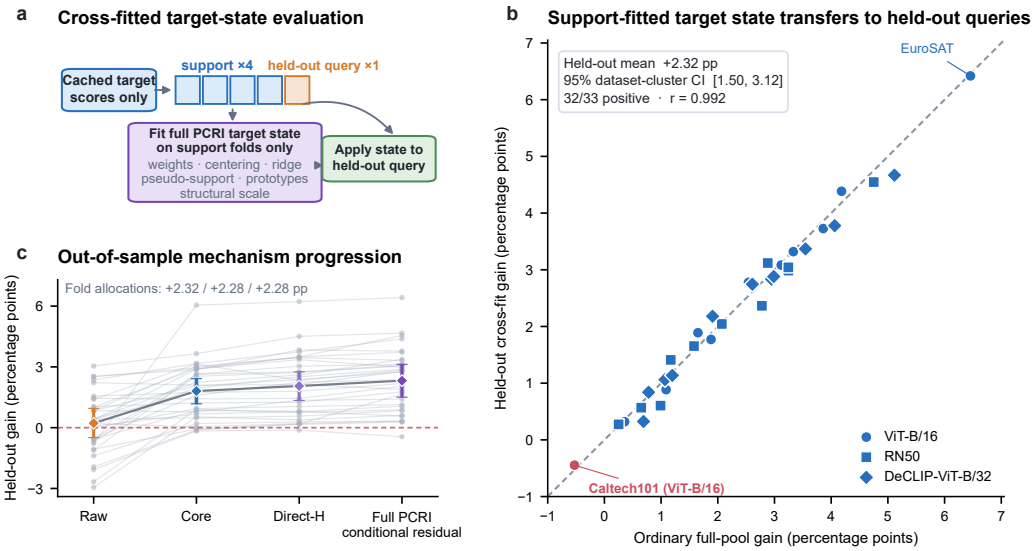


Figure 5: Cross-fitted target-state transfer under strict score-only access. (a) Five-fold query-excluded fitting. (b) Across 33 settings (401,088 held-out predictions), support-fitted gains retain 97.6% of the full-pool gain (32/33 positive; $r = 0.992$); shapes denote backbones. (c) Held-out Raw–Core–Direct-H–Full PCRI trajectories and dataset-cluster intervals. Cross-fitting removes query self-influence, not dependence on other unlabeled samples.

The held-out mechanism stages gain +0.223 pp for raw-score prototypes, +1.808 pp for the Core, +2.056 pp for Direct-H, and +2.320 pp for full PCRI (Figure 5c). The largest step is therefore the shared Core; prompt information adds further gain even when the target state is fitted without the query. These stage means describe a matched progression, not independent causal effects.

The single ViT-B/16 Caltech101 failure remains visible. Appendix F separately retains the structural-support coverage and imbalance boundaries; an exploratory score-only attenuation is not part of canonical PCRI and provides no universal no-harm guarantee.

Generality across representative score-only prompt aggregators. The same frozen PCRI configuration improves MPE, WPE, and CARPRT on ten non-ImageNet CLIP ViT-B/16 datasets:

Table 3: Generality across representative score-only prompt aggregators.

Base predictor	Base acc.	+PCRI acc.	Gain (pp)
MPE	64.176	67.190	+3.014
WPE	64.829	68.401	+3.573
CARPRT	67.299	69.984	+2.685

PCRI is applied after each aggregator, using the same target-state procedure and frozen hyperparameters rather than retuning the correction for a particular prompt-weighting rule. For MPE, WPE, and CARPRT, the respective paired-gain 95% intervals are [+2.002,+3.964], [+2.110,+5.345], and [+1.529,+3.895] pp (Appendix E.1). Each interval excludes zero in this ten-dataset comparison. The corrected accuracies retain the base predictors' ordering: CARPRT remains ahead of WPE and MPE. Thus, the score-space correction complements the starting predictor without erasing differences among prompt aggregators; these results do not establish universality.

The prompt partition serves a different purpose. With $G = 4$, five fixed random partitions yield gains from +2.578 to +2.730 pp, compared with +2.685 pp for the deterministic SHA-256 partition on the same ten datasets (Appendix E.4). Four random partitions improve 9/10 datasets and one improves 10/10; the deterministic partition also improves 9/10. The hash supplies reproducible grouping, not a learned semantic taxonomy. This comparison indicates that the observed gain is not tied to the exact hash membership within the tested setting. Post-score efficiency is reported in Appendix G.

6 LIMITATIONS AND SCOPE

PCRI remains target-transductive: Figure 5 removes query self-influence but fits state from other unlabeled samples. Sparse or unrepresentative support can cause negative transfer (Figure 4; Appendix F); a class absent from base pseudo-labels has no valid structural prototype. PCRI cannot guarantee recovery from severe semantic misalignment, and Proposition F.1 excludes a distribution-free sample-wise no-harm guarantee from cached scores alone. Appendix F.3 reports a development-only attenuation outside canonical PCRI, not a universal safety mechanism.

PCRI also requires prompt-level class scores, prompt identities, and class order; top-1 or truncated-score APIs are outside scope. Efficiency is post-score, and evaluation is limited to zero-shot image classification; detection, segmentation, retrieval, and online continual adaptation remain open.

7 CONCLUSION

Prompt-class scores provide repeated text-induced measurements, not only predictions to aggregate. PCRI recovers shared class-contrast geometry, retains prompt variation conditional on it, and anchors target correction to frozen semantic scores. This shared–conditional decomposition unifies response invariance, conditional complementarity, and semantic anchoring while making its observability boundary explicit. Across 33 matched settings, PCRI improves matched CARPRT by +2.378 pp on average and improves 32/33 settings; query-excluded cross-fitting preserves +2.320 pp, or 97.6% of the full-pool gain. The dominant Core gain is attributable to class-contrast standardization, Core conditioning provides an additional matched improvement, and stress tests restrict the claim to sufficiently representative unlabeled pools rather than universal robustness.

AI USE STATEMENT

Generative AI tools were used in a limited, assistive role for language editing, literature organization, and non-binding feedback on conceptual framing, hypothesis formulation, theoretical arguments, experimental design, implementation, figures, and result presentation. The authors originated the research direction and method, made all scientific decisions, ran and validated the experiments, independently verified the mathematical claims and proofs, and reviewed, revised, and approved the final submission. The authors take full responsibility for its contents.

REPRODUCIBILITY STATEMENT

The paper separates score extraction, score-only adaptation, and label-based evaluation; matched CARPRT and PCRI consume identical frozen score tensors with fixed sample, prompt, and class order. Appendices A and B specify the method, inputs, benchmark protocol, and statistical procedures; Appendices C–F report the setting-level, mechanism, and stress-test results; Appendix G gives post-score complexity; and Appendix H inventories the released implementation and selected result records. The companion code package does not bundle the upstream datasets, model weights, frozen score caches, or every per-trial appendix record; end-to-end reruns require the relevant external assets.

ETHICS STATEMENT

This work evaluates a post-processing method on established image-classification benchmarks and pretrained vision-language models; it does not collect new participant data or conduct interventions with people. Dataset images, model checkpoints, and prompt resources remain subject to their original licenses and terms, which are not superseded by the released code license. PCRI may inherit biases and errors from its base predictors and target data; improvements in classification accuracy do not establish fairness, calibration, or deployment safety. Section 6 and Appendix F report the method's operating limits and observed negative-transfer cases.

REFERENCES

James Urquhart Allingham, Jie Ren, Michael W. Dusenberry, Xiuye Gu, Yin Cui, Dustin Tran, Jeremiah Z. Liu, and Balaji Lakshminarayanan. A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models. In *ICML*, 2023.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101: Mining discriminative components with random forests. In *ECCV*, 2014.

Malik Boudiaf, Imtiaz Masud Ziko, Jérôme Rony, Jose Dolz, Pablo Piantanida, and Ismail Ben Ayed. Information maximization for few-shot learning. In *NeurIPS*, 2020.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Ruijiang Dong, Zesheng Ye, Jianzhong Qi, Lei Feng, Feng Liu, Gang Niu, and Masashi Sugiyama. CARPRT: Class-aware zero-shot prompt reweighting for black-box vision-language models. In *ICLR*, 2026.

Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. In *NeurIPS*, 2024.

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 *CVPR Workshop*, 2004.

Cheng-Ming Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *ICCV*, 2023.

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

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *CVPR*, 2024.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *CVPR*, 2023.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, 2013.

Yangguang Li, Fang Liang, Licheng Zhao, Yufei Cui, Wanli Ouyang, Jing Shao, Feng Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2022.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020.

Yuhang Lu, Jialiang Liu, Yong Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *CVPR*, 2022.

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

Fan'an Meng, Chaoran Cui, Hongjun Dai, and Shuai Gong. Black-box test-time prompt tuning for vision-language models. In *AAAI*, volume 39, pp. 6099–6107, 2025.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

Shuaicheng Niu, Jiahao Wu, Yifan Zhang, Yifan Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *ICML*, 2022.

Shuaicheng Niu, Jiahao Wu, Yifan Zhang, Zhiquan Wen, Yifan Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *ICLR*, 2023.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Jameel Abdul Samadh, Muhammad Haris Gani, Noor Hussein, Muhammad Uzair Khattak, Muzammal Naseer, Fahad Shahbaz Khan, and Salman H. Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *NeurIPS*, 2023.

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 *NeurIPS*, 2022.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. SuS-X: Training-free name-only transfer of vision-language models. In *ICCV*, 2023.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021.

Haoyan Wu, Yahao Liu, Yinjie Lei, Lixin Duan, and Wen Li. Dynamic logits adjustment and exploration for test-time adaptation in vision language models. In *CVPR*, pp. 3143–3153, 2026.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *CVPR*, pp. 23783–23793, 2024.

Maxime Zanella, Clément Fuchs, Christophe De Vleeschouwer, and Ismail Ben Ayed. Realistic test-time adaptation of vision-language models. In *CVPR*, pp. 25103–25112, 2025.

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

Lihua Zhou, Mao Ye, Shuaifeng Li, Nianxin Li, Xiatian Zhu, Lei Deng, Hongbin Liu, and Zhen Lei. Bayesian test-time adaptation for vision-language models. In *CVPR*, pp. 29999–30009, 2025.

Imtiaz Masud Ziko, Jose Dolz, Eric Granger, and Ismail Ben Ayed. Laplacian regularized few-shot learning. In *ICML*, 2020.

SUPPLEMENTARY MATERIAL

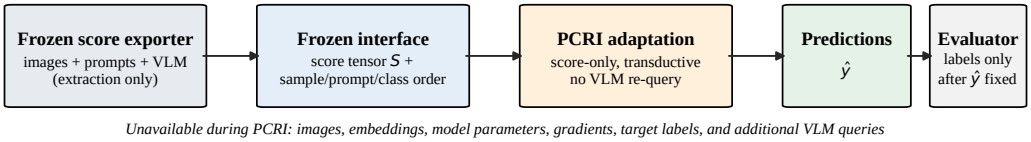
Organization. Appendix A gives the exact PCRI implementation under the shared–conditional score-space principle used in the main paper: shared class-contrast geometry through the Score-space Core, conditional prompt innovation beyond that Core, and semantically anchored target correction. Appendix B details the experimental and statistical protocols, Appendix C reports the complete matched evaluation, Appendix D maps mechanism experiments to their corresponding claims, Appendix E studies generality and operating conditions, Appendix F characterizes stress tests and the exploratory coverage-aware extension, Appendix G reports post-score computational cost, and Appendix H documents reproducibility assets and provenance.

A EXACT METHOD SPECIFICATION AND SCORE-ONLY ACCESS

A.1 OBSERVABLE INPUTS AND ACCESS CONSTRAINTS

PCRI starts only after score extraction has completed. For N target samples, P prompts, and C classes, the observable score tensor is $S \in \mathbb{R}^{N\times P\times C}$. Score extraction may internally use normalized image/text representations and the frozen checkpoint logit scale, but those objects are not accessible to PCRI. During adaptation, PCRI can use the frozen scores, opaque sample order, prompt identities, and class order; it cannot access raw images, embeddings, intermediate activations, model parameters, gradients, target labels, or additional VLM queries.

Figure S1. Score-only execution boundary. Labels become visible only to the evaluator after predictions are fixed; they are not available to PCRI adaptation or inference.



$$S_{n,p,c} = \tau\, \bar{t}_{p,c}^{\mathsf{T}} \bar{z}_n, \qquad \tau = \exp(\gamma), \qquad S \in \mathbb{R}^{N\times P\times C}. \tag{A1}$$

A.2 SHARED CLASS-CONTRAST GEOMETRY AND THE SCORE-SPACE CORE

The base score matrix B is produced by a fixed score-only aggregator. Base scores rank classes, but their sample-dependent common offset and positive scale can obscure relational comparisons across target samples. For the weighted prompt aggregators used in the reported experiments, once target-level prompt weights are fixed, define the effective text measurement direction $m_c = \sum_p w_{p,c}\bar{t}_{p,c}$ and $M = [m_1, \ldots, m_C]$. Then $b_n = \tau M^{\mathsf{T}} \bar{z}_n$, so the aggregated class-score vector can be interpreted as an observable set of text-induced measurements. PCRI does not assume an orthogonal text basis and does not reconstruct $\bar{z}_n$.

$$\begin{aligned} m_c &= \sum_{p=1}^{P} w_{p,c}\, \bar{t}_{p,c}, \\ M &= [m_1, \ldots, m_C], \\ b_n &= \tau M^{\mathsf{T}} \bar{z}_n. \end{aligned} \tag{A2}$$

$$P_C = I - \frac{1}{C}\mathbf{1}\mathbf{1}^{\mathsf{T}}, \qquad (P_C b_i)^{\mathsf{T}}(P_C b_j) = \tau^2 \bar{z}_i^{\mathsf{T}} M P_C M^{\mathsf{T}} \bar{z}_j. \tag{A3}$$

The identity above motivates a surrogate response geometry but does not assert equivalence to the inaccessible visual feature space. The main paper summarizes the following two properties as Proposition 1.

Proof of Proposition 1. First, P_C is the orthogonal projector onto the class-centered subspace. For any vector v,

$$v^\top M P_C M^\top v = (M^\top v)^\top P_C (M^\top v) = \|P_C M^\top v\|_2^2 \geq 0. \tag{A4}$$

so $K_M = M P_C M^\top$ is positive semidefinite, and the centered score inner product is a text-induced positive-semidefinite bilinear similarity.

Second, on the non-degenerate domain $\mathcal{X} = \{b \in \mathbb{R}^C : P_C b \neq 0\}$, define $q(b) = P_C b / \|P_C b\|_2$. If $b' = ab + \delta \mathbf{1}$ with $a > 0$, then $P_C b' = a P_C b$ and therefore $q(b') = q(b)$. Conversely, if $q(b') = q(b)$, set $a = \|P_C b'\|_2 / \|P_C b\|_2 > 0$. Then $P_C(b' - ab) = 0$, so $b' - ab$ lies in the null space of P_C, which is $\mathrm{span}\{\mathbf{1}\}$. Hence there is a $\delta \in \mathbb{R}$ such that $b' = ab + \delta \mathbf{1}$. This proves that q is a maximal invariant for sample-wise positive-affine transformations. With the population mean and standard deviation,

$$\frac{b'_n - \mu'_n \mathbf{1}}{\sigma'_n} = \frac{b_n - \mu_n \mathbf{1}}{\sigma_n} = \sqrt{C}\, \frac{P_C b_n}{\|P_C b_n\|_2}. \tag{A5}$$

Thus the idealized standardized row $U_0^\circ(b) = \sqrt{C}\, q(b)$ retains exactly the information distinguishable modulo this nuisance group. Equivalently, every statistic T that is invariant on $\mathcal{X}$ factors through q: defining $\widetilde{T}(q(b)) = T(b)$ is well posed because equal values of q identify the same group orbit. Constant score rows lie outside $\mathcal{X}$ and are handled by the numerical floor in Eq. (A6). The implemented U_0 equals U_0° whenever that floor is inactive. The subsequent dataset-level BlockNorm remains part of the operational Core, but is not itself claimed to be maximal.

Corollary A.1 (Conditional bounded distortion of centered measurements). Define the centered text-induced measurement operator $A_M = P_C M^\top$ and $\phi_M(\bar{z}) = \tau A_M \bar{z}$. Then

$$\langle \phi_M(\bar{z}_i), \phi_M(\bar{z}_j) \rangle = \tau^2 \bar{z}_i^\top A_M^\top A_M \bar{z}_j = \tau^2 \bar{z}_i^\top K_M \bar{z}_j.$$

Let $\mathcal{V}$ be any task-relevant difference subspace. If the restricted operator $A_M|_{\mathcal{V}}$ is non-degenerate and all of its singular values lie in $[s_{\min}, s_{\max}]$, with $0 < s_{\min} \leq s_{\max} < \infty$, then for any $\bar{z}_i - \bar{z}_j \in \mathcal{V}$,

$$\tau s_{\min} \|\bar{z}_i - \bar{z}_j\|_2 \leq \|\phi_M(\bar{z}_i) - \phi_M(\bar{z}_j)\|_2 \leq \tau s_{\max} \|\bar{z}_i - \bar{z}_j\|_2.$$

This follows directly from the singular-value bounds applied to $A_M(\bar{z}_i - \bar{z}_j)$, together with $P_C(b_i - b_j) = \tau A_M(\bar{z}_i - \bar{z}_j)$. The statement is conditional: it does not assert that a given prompt bank satisfies the restricted-spectrum condition. It applies to the centered linear measurements before per-sample standardization and BlockNorm, and therefore does not by itself establish a property of the later prototype or fusion stages or guarantee improved classification.

These properties justify the representation interpretation used in the main paper; they do not imply that the induced geometry recovers the hidden VLM feature space or guarantees improved classification. The exact safeguards below cover degenerate rows.

$$\begin{aligned} U_0[n,c] &= \frac{B_{n,c} - \mu_n}{\max(\sigma_n, 10^{-6})}, \\ \mu_n &= \frac{1}{C}\sum_j B_{n,j}, \qquad \sigma_n^2 = \frac{1}{C}\sum_j (B_{n,j} - \mu_n)^2. \end{aligned} \tag{A6}$$

$$\begin{aligned} \bar{X} &= \frac{1}{N}\sum_i X_{i,:}, \\ \mathrm{BlockNorm}(X)_{n,:} &= \frac{X_{n,:} - \bar{X}}{\max(\|X_{n,:} - \bar{X}\|_2, 10^{-12})}, \qquad U = \mathrm{BlockNorm}(U_0). \end{aligned} \tag{A7}$$

A.3 Conditional prompt innovation beyond the Core

Aggregation from the $P \times C$ prompt tensor to C base scores suppresses prompt-conditioned variation across textual measurement views. PCRI therefore standardizes each sample-prompt score

vector across classes, removes the sample-wise prompt mean, groups the signed deviations deterministically, and flattens them to H. The complete UTF-8 prompt string is hashed with SHA-256; the first eight digest bytes are interpreted as an unsigned big-endian integer and reduced modulo $G = 4$. The hash is intentionally a deterministic, order-invariant, non-semantic partition rather than a semantic clustering rule or a learned prompt-quality estimator. Its role is reproducible compression of prompt deviations without introducing additional learned parameters or label-dependent grouping. Appendix E.4 shows that five fixed random partitions yield similar performance, so the method does not rely on a privileged semantic interpretation of SHA-256; we make no Johnson–Lindenstrauss, orthogonality, or distance-preservation claim for this grouping rule.

$$\begin{aligned} Z_{n,p,c} &= \frac{S_{n,p,c} - \mu^S_{n,p}}{\max(\sigma^S_{n,p}, 10^{-6})}, \\ R_{n,p,c} &= Z_{n,p,c} - \frac{1}{P}\sum_q Z_{n,q,c}. \end{aligned} \tag{A8}$$

$$\begin{aligned} g(p) &= \mathrm{Int}_{\mathrm{BE}}(\mathrm{SHA256}(\mathrm{UTF8}(t_p))[0{:}8]) \bmod G, \qquad G = 4, \\ \mathcal{H}_{n,g,c} &= \frac{1}{|\mathcal{P}_g|}\sum_{p\in\mathcal{P}_g} R_{n,p,c}, \\ H_n &= [\mathcal{H}_{n,0,:}; \ldots; \mathcal{H}_{n,G-1,:}] \in \mathbb{R}^{GC}. \end{aligned} \tag{A9}$$

Prompt-level responses contain both variation already predictable from the Score-space Core and potentially complementary variation suppressed by aggregation. The implementation first places the grouped descriptor in the same normalization regime as the Core, $\bar{H} = \mathrm{BlockNorm}(H)$, and then ridge-residualizes this normalized view against U. The complete normalized ridge-residual construction is given in Eq. (A10). Thus the pre-ridge and post-residual BlockNorm operations are both part of the frozen implementation. Because $\lambda > 0$, the residual is not an exact Euclidean orthogonal projection and does not imply statistical independence. We interpret H_{res} as conditional prompt innovation - normalized prompt-conditioned variation that is less linearly predictable from the Score-space Core. The Core and residual views are concatenated and normalized once more before prototype reasoning: $F = \mathrm{BlockNorm}([U; \beta H_{\mathrm{res}}])$, with $\lambda = 30$ and $\beta = 0.75$ fixed from the development setting.

$$\begin{aligned} \bar{H} &= \mathrm{BlockNorm}(H), & A^* &= (U^\top U + \lambda I)^{-1} U^\top \bar{H}, & \lambda &= 30, \\ H_{\mathrm{res}} &= \mathrm{BlockNorm}(\bar{H} - UA^*), & F &= \mathrm{BlockNorm}([U; \beta H_{\mathrm{res}}]), & \beta &= 0.75. \end{aligned} \tag{A10}$$

Equivalently, with the ridge shrinkage operator

$$\begin{aligned} \Pi_\lambda(U) &= U(U^\top U + \lambda I)^{-1} U^\top, \\ \bar{H} - UA^* &= (I - \Pi_\lambda(U))\bar{H}, \qquad U^\top(\bar{H} - UA^*) = \lambda A^*. \end{aligned} \tag{A11}$$

The final equality is the ridge normal equation. These identities formalize the conditioning interpretation: the residual removes the component of the normalized grouped prompt view predicted linearly from the Core under the frozen ridge penalty. For $\lambda > 0$, $\Pi_\lambda(U)$ is generally not an idempotent orthogonal projector, so neither exact Euclidean orthogonality nor statistical independence is claimed.

A.4 Semantically anchored target correction

Relational geometry alone is equivariant to a simultaneous permutation of its unlabeled prototype identities and therefore does not identify class semantics. Frozen base scores break this class-permutation symmetry: pseudo-labels and confidence come directly from the frozen base scores. Confidence is computed as the maximum class probability after applying softmax to the checkpoint-scaled base logits B; PCRI applies no additional temperature. Within each nonempty pseudo-class, samples at or above the PyTorch median confidence are retained as support. This is a class-adaptive, parameter-free rank split: apart from ties, any strictly increasing transformation of confidence values

within one pseudo-class preserves the selected ordering, and every nonempty pseudo-class retains at least one support sample. Selected sample vectors are individually L2-normalized, averaged, and the resulting prototype is L2-normalized. Structural scores are cosine similarities to valid prototypes. In the final fusion, the base predictor remains the semantic anchor and the prototype branch acts only as a target-conditioned structural correction. Scale matching aligns global numerical dispersion so the fixed fusion weight is meaningful; it does not introduce new class evidence or replace the base semantic ranking.

$$\begin{aligned} \tilde{y}_n &= \arg\max_c B_{n,c}, & r_n &= \max_c [\mathrm{softmax}(B_{n,:})]_c, \\ \mathcal{I}_c &= \{n : \tilde{y}_n = c\}, & r_c^{\mathrm{med}} &= \mathrm{median}(\{r_i : i \in \mathcal{I}_c\}), \\ \mathcal{I}_c^+ &= \{n \in \mathcal{I}_c : r_n \geq r_c^{\mathrm{med}}\}. \end{aligned} \tag{A12}$$

$$\begin{aligned} x_n^\circ &= \frac{X_{n,:}}{\max(\|X_{n,:}\|_2, 10^{-12})}, & q_c &= \frac{1}{|\mathcal{I}_c^+|} \sum_{n \in \mathcal{I}_c^+} x_n^\circ, \\ p_c &= \frac{q_c}{\max(\|q_c\|_2, 10^{-12})}, & [G_{\mathrm{struct}}(X)]_{n,c} &= (x_n^\circ)^\mathsf{T} p_c. \end{aligned} \tag{A13}$$

$$\begin{aligned} \kappa_F &= \frac{\mathrm{Std}(B)}{\max(\mathrm{Std}([G_{\mathrm{struct}}(F)]_{\mathrm{finite}}), 10^{-12})}, \\ \widetilde{G}_{\mathrm{struct}} &= \kappa_F G_{\mathrm{struct}}(F), \\ \Delta(F; B) &= \widetilde{G}_{\mathrm{struct}} - B, \\ B_{\mathrm{PCRI}} &= B + \alpha\Delta(F; B) = (1-\alpha)B + \alpha\widetilde{G}_{\mathrm{struct}}, \qquad \alpha = 0.50, \\ \hat{y}_n &= \arg\max_c [B_{\mathrm{PCRI}}]_{n,c}. \end{aligned} \tag{A14}$$

When the finite structural standard deviation exceeds the numerical floor, $\mathrm{Std}([\widetilde{G}_{\mathrm{struct}}]_{\mathrm{finite}}) = \mathrm{Std}(B)$. Thus the fixed α mixes comparably dispersed score fields rather than uncontrolled raw numerical units. The equivalent correction form $B_{\mathrm{PCRI}} = B + \alpha\Delta$ also makes explicit that the structural branch is anchored to the base semantic predictor instead of replacing it with an independently trained classifier.

A.5 Degenerate and empty-support cases

If no sample receives pseudo-label c, class c is invalid and no fallback prototype is constructed. Its structural entries are excluded from the finite-score standard deviation used for scale matching and are materialized as a large negative structural score after scaling. For any nonempty target pool, at least one pseudo-class is nonempty, and the class-wise median rule retains at least one support sample in every nonempty pseudo-class; hence the evaluated protocol cannot produce an all-invalid prototype set. Invalid structural classes therefore cannot win the structural branch; PCRI does not silently replace them with a label-informed or feature-based fallback. Thus this branch cannot rescue a class missing from base pseudo-label support; Appendix F tests the resulting coverage boundary.

A.6 COMPLETE INFERENCE ALGORITHM

Algorithm 1 Complete PCRI inference under the frozen score-only interface. Shared class contrasts form the Core; grouped prompt responses are conditioned on that Core; the joint representation supplies target relations while frozen scores anchor the final correction.

Require: Cached scores $S \in \mathbb{R}^{N\times P\times C}$; fixed $\mathcal{A}_{\text{base}}$; $\alpha = 0.50$, $\beta = 0.75$, $\lambda = 30$, $G = 4$, $\epsilon_{\text{std}} = 10^{-6}$, $\epsilon_{\text{BN}} = 10^{-12}$.

$B \leftarrow \mathcal{A}_{\text{base}}(S)$

$U \leftarrow \text{BlockNorm}(z_{\text{class}}(B))$

$\tilde{y}_n \leftarrow \arg\max_c B_{n,c}$; $r_n \leftarrow \max_c[\text{softmax}(B_{n,:})]_c$

$\mathcal{S}_c \leftarrow \{n : \tilde{y}_n = c,\ r_n \geq \text{median}(\{r_j : \tilde{y}_j = c\})\}$

$Z_{n,p,c} \leftarrow (S_{n,p,c} - \mu^S_{n,p})/\max(\sigma^S_{n,p}, \epsilon_{\text{std}})$

$R_{n,p,c} \leftarrow Z_{n,p,c} - P^{-1}\sum_{q=1}^{P} Z_{n,q,c}$

$\mathcal{H}_{n,g,c} \leftarrow |\mathcal{P}_g|^{-1}\sum_{p\in\mathcal{P}_g} R_{n,p,c}$; $H_n \leftarrow \text{vec}(\mathcal{H}_{n,0,:}, \ldots, \mathcal{H}_{n,G-1,:})$

$\bar{H} \leftarrow \text{BlockNorm}(H)$

$A^* \leftarrow (U^\top U + \lambda I)^{-1}U^\top\bar{H}$; $H_{\text{res}} \leftarrow \text{BlockNorm}(\bar{H} - UA^*)$

$F \leftarrow \text{BlockNorm}([U; \beta H_{\text{res}}])$

$p_c \leftarrow \text{normalize}(|\mathcal{S}_c|^{-1}\sum_{n\in\mathcal{S}_c}\text{normalize}(F_n))$ for valid c

$[G_{\text{struct}}]_{n,c} \leftarrow \cos(F_n, p_c)$; invalid classes receive $-\infty$

$\kappa \leftarrow \text{Std}(B)/\max(\text{Std}(G_{\text{struct,finite}}), 10^{-12})$; $B_{\text{PCRI}} \leftarrow (1-\alpha)B + \alpha\kappa G_{\text{struct}}$

$\hat{y}_n \leftarrow \arg\max_c[B_{\text{PCRI}}]_{n,c}$, $n = 1, \ldots, N$

Ensure: Final predictions $\hat{y} \in \{1, \ldots, C\}^N$.

B EXPERIMENTAL AND STATISTICAL PROTOCOLS

B.1 DATASETS AND SAMPLE COUNTS

The main matched evaluation contains 11 datasets and three backbones (33 settings). The benchmarks are Caltech101 (Fei-Fei et al., 2004), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), FGVC Aircraft (Maji et al., 2013), Food101 (Bossard et al., 2014), Flowers102 (Nilsback & Zisserman, 2008), Oxford-IIIT Pets (Parkhi et al., 2012), Stanford Cars (Krause et al., 2013), SUN397 (Xiao et al., 2010), UCF101 (Soomro et al., 2012), and ImageNet (Deng et al., 2009). Table S1 lists the target-pool size used for each dataset.

Table S1. Evaluation target-pool sizes used by all three backbones.

Dataset	Target samples
Caltech101	2,465
DTD	1,692
EuroSAT	8,100
FGVC Aircraft	3,333
Food101	30,300
Flowers102	2,463
Oxford-IIIT Pets	3,669
Stanford Cars	8,041
SUN397	19,850
UCF101	3,783
ImageNet	50,000

B.2 BACKBONES, CHECKPOINTS, PREPROCESSING, AND TOKENIZATION

Table S2. Frozen checkpoints, preprocessing, and tokenization for CLIP (Radford et al., 2021) and DeCLIP (Li et al., 2022). Hashes are abbreviated here; Appendix H reports the full SHA-256 identifiers and source notes.

Backbone	Checkpoint SHA-256	Image preprocessing	Tokenization
OpenAI CLIP ViT-B/16	5806E77C…DF416F	Resize 224 bicubic; center crop 224; CLIP normalization	SimpleTokenizer; context 77
OpenAI CLIP RN50	AFEB0E10…B6762	224 input resolution; same CLIP normalization	SimpleTokenizer; context 77
DeCLIP ViT-B/32	168D4CA6…8BC3E	Resize 256 bicubic; center crop 224; CLIP normalization	DeCLIP SimpleTokenizer with CLIP BPE vocabulary; context 77

B.3 Prompt bank and deterministic grouping

The frozen prompt bank contains 247 templates and is byte-identical to the matched CARPRT runtime copy (full file SHA-256 in Table S18). Under $G = 4$, group counts are [68, 47, 67, 65]. PCRI neither generates nor modifies prompts. The companion artifact provides the ordered 247-row index, prompt-string SHA-256 hashes, and deterministic group IDs (Table S19), but not template text. The text comes from the frozen upstream prompt bank; Appendix A.3 specifies the hashing rule.

B.4 Development and evaluation separation

Method design and hyperparameter selection are restricted to CIFAR-100 with CLIP ViT-B/16, and the 11 evaluation datasets receive no dataset-specific tuning. The implementation separates score extraction, score-only adaptation, and label-based evaluation. Matched CARPRT and PCRI receive the same frozen score tensor with identical sample, prompt, and class order. Published CARPRT values are used only as literature references; all paired PCRI comparisons use the matched CARPRT predictions.

B.5 Controlled sampling protocols

Table S3. Controlled subset construction. Ground-truth labels are used only by the evaluation harness to construct class-coverage/imbalance stress-test subsets and to compute metrics; PCRI never receives them.

Analysis	Scope	Conditions	Seeds	Replicates	Construction
Adaptation-support fraction	10 non-ImageNet datasets, ViT-B/16	10/25/50/75/100%	20260841–20260843	30 trials/fraction	Frozen base score field; uniform sample permutation; $\mathrm{round}(fN)$ PCRI support
Prompt budget	10 non-ImageNet datasets, ViT-B/16	16/32/64/128/247	20260841–20260843	30 trials/budget	Nested prefixes of one seeded permutation; selected indices sorted; hash groups recomputed
Cross-fitted target state	11 datasets × 3 backbones	5 folds	20260913–20260915	33 settings/seed	Every query excluded from weights, centering, ridge, pseudo-support, prototypes, and scale; labels opened only after prediction
Structural-support class coverage	Caltech101 / Food101 / SUN397 / ImageNet, ViT-B/16	5/10/25/50/100%	20260841–20260845	20 trials/fraction	Frozen base score field; selected-class samples supply PCRI support
Class imbalance	Same four datasets, ViT-B/16	balanced; Dirichlet 1/.5/.1	20260841–20260845	20 trials/condition	$\mathrm{round}(0.5N)$; class-weighted sampling without replacement

B.6 Bootstrap and paired statistical analysis

The main-result interval uses 100,000 bootstrap replicates with RNG seed 20260825. The resampling unit is dataset: each sampled dataset carries all three backbone results together. Each replicate draws 11 datasets with replacement and computes the setting-first macro mean over the resulting 33 paired differences. The reported 95% interval is the two-sided percentile interval at the 2.5th and 97.5th percentiles, not BCa. This protocol yields [+1.554,+3.196] pp for the overall PCRI gain. All 11 three-backbone dataset means are positive; leave-one-dataset-out means span +2.147 to +2.588 pp, and an exact two-sided sign test gives $p = 2/2^{11} = 0.00098$. These are benchmark-suite sensitivity summaries, not population-generalization guarantees.

For the secondary mechanism and target-composition analyses, we report the final 95% confidence intervals as uncertainty summaries and avoid stronger claims that depend on additional resampling assumptions. Appendix H.5 summarizes the level of statistical metadata available for these analyses.

C Complete Main Results

C.1 Full 33-setting matched results

Table S4 reports the matched CARPRT and PCRI absolute accuracies for all 33 settings. Every paired gain is computed against matched CARPRT using the identical frozen score tensor; paper-reported CARPRT values are not used for paired differences. Gains are computed from unrounded accuracies; displayed accuracies and gains are rounded independently.

Table S4. Complete matched evaluation across 11 datasets and three backbones, reported separately for each backbone.

(a) CLIP ViT-B/16

Dataset	N	Matched CARPRT (%)	PCRI (%)	Gain (pp)
Caltech101	2,465	94.523	93.996	-0.527
DTD	1,692	48.936	51.478	+2.541
EuroSAT	8,100	55.049	61.506	+6.457
FGVC Aircraft	3,333	24.632	26.283	+1.650
Food101	30,300	85.861	86.218	+0.356
Flowers102	2,463	71.376	75.558	+4.182
Oxford-IIIT Pets	3,669	89.452	91.333	+1.881
Stanford Cars	8,041	66.173	69.506	+3.333
SUN397	19,850	66.937	70.055	+3.118
UCF101	3,783	70.050	73.910	+3.859
ImageNet	50,000	68.792	69.880	+1.088

(b) CLIP ResNet-50

Dataset	N	Matched CARPRT (%)	PCRI (%)	Gain (pp)
Caltech101	2,465	88.600	89.249	+0.649
DTD	1,692	40.662	43.440	+2.778
EuroSAT	8,100	36.951	38.531	+1.580
FGVC Aircraft	3,333	16.832	18.002	+1.170
Food101	30,300	76.706	76.957	+0.251
Flowers102	2,463	65.408	70.158	+4.750
Oxford-IIIT Pets	3,669	85.991	88.062	+2.071
Stanford Cars	8,041	56.374	59.619	+3.246
SUN397	19,850	61.244	64.489	+3.244
UCF101	3,783	63.310	66.191	+2.881
ImageNet	50,000	60.398	61.388	+0.990

(c) DeCLIP ViT-B/32

Dataset	N	Matched CARPRT (%)	PCRI (%)	Gain (pp)
Caltech101	2,465	93.996	94.686	+0.690
DTD	1,692	42.967	46.513	+3.546
EuroSAT	8,100	38.815	42.877	+4.062
FGVC Aircraft	3,333	7.741	8.521	+0.780
Food101	30,300	72.931	74.000	+1.069
Flowers102	2,463	81.608	86.724	+5.116
Oxford-IIIT Pets	3,669	81.875	83.783	+1.908
Stanford Cars	8,041	52.195	55.142	+2.947
SUN397	19,850	63.940	66.549	+2.610
UCF101	3,783	60.005	62.992	+2.987
ImageNet	50,000	66.310	67.508	+1.198

Table S5. Aggregate paired statistics over the complete matched evaluation.

Scope	Matched (%)	PCRI (%)	Mean gain	Median gain	W/T/L	95% CI
All 33 settings	**62.322**	**64.700**	**+2.378**	**+2.541**	**32/0/1**	**[+1.554,+3.196]**
CLIP ViT-B/16	67.435	69.975	+2.540	+2.541	10/0/1	–
CLIP ResNet-50	59.316	61.462	+2.146	+2.071	11/0/0	–
DeCLIP ViT-B/32	60.217	62.663	+2.447	+2.610	11/0/0	–

Visual summary of the complete evaluation. Figure S3 complements the exact numerical tables by separating two views of the result structure. Panel (a) visualizes the full main-comparison accuracy matrix across methods, datasets, and backbones. Panel (b) isolates the exact PCRI gain over the matched CARPRT prediction for each dataset–backbone setting, using the same frozen score tensors as Table S4. The heatmap makes the setting-level consistency visible without replacing the numerical tables.

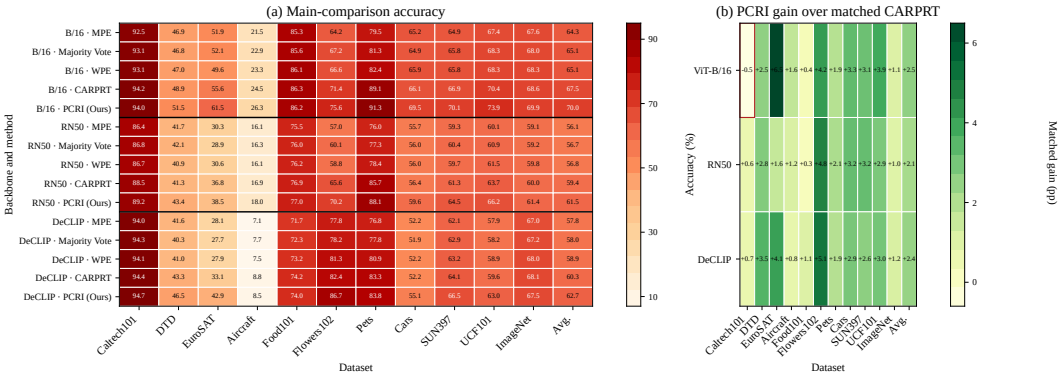


Figure S3. Visual summary of the complete benchmark results. (a) Absolute classification accuracies from the main comparison, with PCRI rows visually outlined. (b) Exact paired PCRI gain over matched CARPRT for every dataset–backbone setting; the rightmost column reports the backbone-wise mean gain. Cell values are percentage points. This figure is a visualization of already reported results and does not introduce an additional evaluation.

C.2 Setting-level consistency

PCRI improves 32 of 33 matched settings. The only decrease is CLIP ViT-B/16 on Caltech101 (94.523%→93.996%, −0.527 pp). ImageNet improves for all three backbones, with gains of +1.088, +0.990, and +1.198 pp.

D Mechanism Analysis

D.1 Staged ablation and conditional-innovation controls

Table S6. Mechanism, direct-prompt, correspondence, and semantic-anchor controls. Panel A reports mean paired gains over matched CARPRT for successive representation constructions. Panel B reports paired differences that isolate Core conditioning, correspondence, and the frozen-base anchor. Direct-H sets $F_{\text{direct}} = \text{BlockNorm}([U; \beta\bar{H}])$; full PCRI ridge-residualizes $\bar{H}$ against U and applies the same final joint BlockNorm. All remaining components and frozen hyperparameters are unchanged.

Panel A: representation construction	Mean gain vs CARPRT (pp)
CARPRT	0.000
Raw-score prototype	+0.364
Cross-class standardized U_0	+1.844
Score-space Core U	+1.847
Core + direct prompt view (Direct-H)	+2.072
Full PCRI	**+2.378**

Panel B: paired control	Mean Δ	95% CI	W/T/L	ViT-B/16	RN50	DeCLIP
Full PCRI − Direct-H	**+0.306**	**[+0.155,+0.454]**	**28/0/5**	+0.209	+0.357	+0.350
Correct − shuffled prompt view	+1.227	[0.777,1.689]	–	–	–	–
True-Core − permuted-Core	+0.228	[0.028,0.409]	–	–	–	–
Anchored fusion − structural-only	**+1.027**	**[+0.241,+1.847]**	**27/0/6**	+1.248	+1.621	+0.211

The controls separate the shared, conditional, and anchoring claims. Under the matched 33-setting Core decomposition, raw-score prototype reasoning gives +0.364 pp (95% CI [-0.292,+1.076], 21/0/12), cross-class standardized U_0 gives +1.844 pp ([+1.235,+2.439], 31/0/2), and the full Core U gives +1.847 pp ([+1.233,+2.446], 30/0/3). More directly, U_0 minus Raw is +1.480 pp ([+0.946,+2.021], 31/0/2), whereas Full Core minus U_0 is +0.003 pp ([-0.015,+0.027], 13/7/13). Thus almost all of the Core-only mean accuracy gain is already present after cross-class standardization; the subsequent BlockNorm adds negligible aggregate top-1 improvement at this endpoint while remaining part of PCRI's frozen normalization regime. Direct-H reaches +2.072 pp and full PCRI reaches +2.378 pp; their +0.306 pp difference (dataset-clustered 95% CI [+0.155,+0.454]) isolates the additional value of Core conditioning beyond adding the same prompt view. Canonical anchored fusion ($\alpha = 0.5$) exceeds structural-only prediction ($\alpha = 1$) by +1.027 pp across 33 settings, with 27/33 wins and a dataset-clustered 95% CI [+0.241,+1.847]. The shuffled-view and permuted-Core controls test dependence on correct sample/Core correspondence rather than generic extra dimensions or ridge regularization.

Exact construction of correspondence controls. For the within-pseudo-class shuffle control, the true Core and fitted ridge residual are constructed normally; for seeds 20260841–20260843, residual rows are then permuted only among samples sharing the same frozen base pseudo-label before the final joint BlockNorm, prototypes, scale matching, and fusion are recomputed. The pseudo-labels remain fixed. For the permuted-Core control, evaluated on the ten non-ImageNet CLIP ViT-B/16 datasets with the same three seeds, the rows of U are permuted only when fitting the ridge map and forming the residual; that residual is then concatenated with the true Core before the final joint BlockNorm and the otherwise unchanged structural pipeline. Secondary dataset-clustered percentile intervals use 20,000 bootstrap replicates with RNG seed 20260823.

D.2 Cross-fitted conditional predictability

A label-free five-fold diagnostic asks whether the final residual view remains less linearly predictable from the Core after post-residual normalization. On the ten non-ImageNet CLIP ViT-B/16 datasets, folds are fixed by seed 20260823; ridge predictors with the frozen $\lambda = 30$ are fitted on four folds and evaluated on the held-out fold. Macro mean held-out R^2 decreases from 0.616 for $\bar{H}$ to 0.141 for H_{res}, a reduction of 0.474, with a decrease on all ten datasets. This provides out-of-sample representation-level evidence for conditional complementarity; it does not imply statistical independence.

Dataset	$R^2(U \to \bar{H})$	$R^2(U \to H_{\mathrm{res}})$	ΔR^2
Caltech101	0.528	0.181	-0.347
DTD	0.473	0.134	-0.339
EuroSAT	0.484	0.016	-0.469
FGVC Aircraft	0.697	0.119	-0.578
Food101	0.705	0.134	-0.571
Flowers102	0.685	0.223	-0.463
Oxford-IIIT Pets	0.637	0.096	-0.541
Stanford Cars	0.650	0.130	-0.519
SUN397	0.741	0.215	-0.526
UCF101	0.557	0.166	-0.391
Macro mean	**0.616**	**0.141**	**-0.474**

Compact mechanism view. Figure S4 retains the compact visualization of the raw-score/Core/full stages and the two correspondence controls, with the Raw bar updated to the matched 33-setting value. The U_0-only Core decomposition and the additional Direct-H conditioning control are reported numerically in Table S6 rather than expanding the compact figure.

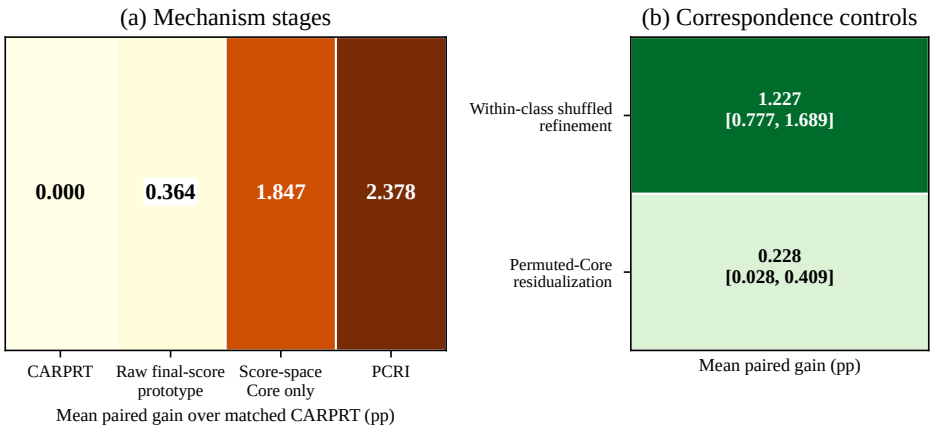


Figure S4. Compact visual summary of the raw-score/Core/full stages and the two original correspondence controls. (a) Mean paired gain for matched CARPRT, the matched 33-setting raw final-score prototype control, the full Score-space Core, and full PCRI. (b) Mean paired differences for the within-class shuffled-view and permuted-Core residualization controls; bracketed values are 95% confidence intervals. The U_0-only decomposition, Direct-H control, and Full-minus-Direct-H statistics are reported in Table S6.

D.3 Score-space geometry diagnostics

On ten non-ImageNet CLIP ViT-B/16 datasets, we test whether the Score-space Core is more separable than raw final scores. It is label-assisted and post-hoc: the score-derived representation is fixed before labels are consulted, and none of these diagnostics enters PCRI adaptation or inference. Table S7 reports per-dataset cross-fitted nearest-prototype (CF-NP) accuracy and the inter-/intra-class separation ratio. The diagnostics therefore test representation suitability for structural reasoning rather than claiming recovery of the hidden VLM embedding. Additional reproducibility notes are summarized in Appendix H.5.

Figure 1 visualization protocol. Figure 1 is a post-hoc, label-assisted visualization from the frozen Food101/OpenAI CLIP ViT-B/16 score cache. The displayed classes are `apple_pie`, `pizza`, and `sushi`; 100 samples per class are drawn without replacement with seed 20260903, and the same 300 indices are used throughout. Figure 1(a) applies separate deterministic 2D PCA fits to matched B and U, so it compares within-representation organization rather than pointwise motion in a shared plane; no individual point is moved, removed, jittered, or relabeled. The three real photographs reuse the previously selected median-margin, correctly classified representative of each displayed class, with center-square cropping only. This illustrative selection does not filter the 300 plotted samples or their pairwise statistics and never enters adaptation.

For Figure 1(b), cosine similarities are computed directly from the corresponding 101-dimensional rows of B or U, not from PCA coordinates. All unordered non-self pairs are retained: 14,850 same-class and 30,000 different-class pairs. Each representation uses 32 equal-width bins with common edges for both relations; density equals the bin count divided by its relation-specific pair count and bin width. The raw-score panel uses $[0.95, 1]$ and the Core panel uses $[-1, 1]$; all observed pairs fall within their displayed range. The empirical, unbinned same-class-minus-different-class mean cosine gap is 0.014 for B and 0.828 for U. No KDE, smoothing, response-coordinate ordering, or artificial jitter is used. Pairwise observations share samples and are not independent replicates. Labels enter only this post-hoc diagnostic. Centering itself changes cosine angles, so the illustrated gap is not an accuracy gain or proof of general class separability.

Table S7. Per-dataset score-space geometry. CF-NP denotes cross-fitted nearest-prototype accuracy.

Dataset	Raw CF-NP	Core CF-NP	Raw inter/intra	Core inter/intra	△ CF-NP	△ ratio
Caltech101	83.448	93.550	5.488	7.992	+10.101	+2.504
DTD	47.163	62.825	2.893	3.837	+15.662	+0.944
EuroSAT	56.383	66.420	3.282	4.466	+10.037	+1.184
FGVC Aircraft	34.233	38.794	6.497	7.407	+4.560	+0.910
Food101	74.858	83.083	4.372	5.374	+8.224	+1.002
Flowers102	87.698	91.555	7.240	9.029	+3.857	+1.789
Oxford-IIIT Pets	90.597	92.805	12.507	18.316	+2.208	+5.809
Stanford Cars	61.846	76.533	7.852	11.966	+14.687	+4.114
SUN397	52.307	72.141	4.184	6.038	+19.834	+1.854
UCF101	67.380	82.924	6.130	7.455	+15.543	+1.325
Macro mean	**65.591**	**76.063**	**6.044**	**8.188**	**+10.472**	**+2.144**

Both diagnostics improve on all 10 datasets for the full Core U. Separately, the matched Core decomposition in Table S6 shows that $U_0 = z_{\text{class}}(B)$ already captures essentially all of the Core-only accuracy gain. Together, these results support the narrower interpretation that cross-class standardization is the performance-critical first transformation, while the geometry diagnostics here compare the full Core U against raw score space. They do not establish equivalence between the Score-space Core and the original visual feature geometry.

D.4 Difficulty localization underlying results

Samples are binned into quintiles within each dataset-backbone setting before setting-level gains are macro-averaged across all 33 settings. Q1 is the lowest-margin/lowest-confidence bin for margin and confidence and the highest-disagreement bin for prompt-group disagreement. Margin and confidence provide the strongest difficulty localization, supporting the claim that PCRI primarily helps uncertain base predictions. Prompt-group disagreement provides a weaker but consistently positive stratification signal and is therefore interpreted more cautiously.

Table S8. Exact mean paired gains (pp) underlying the difficulty-localization figure in the main paper.

Criterion	Q1	Q2	Q3	Q4	Q5
Prediction margin	+7.526	+3.091	+0.997	+0.252	+0.018
Prediction confidence	+6.637	+3.315	+1.475	+0.344	+0.114
Prompt-group disagreement	+2.955	+2.624	+2.499	+1.953	+1.857

E Generality and Operating Conditions

E.1 Generality across base predictors

PCRI uses the same fixed configuration with MPE, WPE, and CARPRT on ten non-ImageNet datasets with CLIP ViT-B/16. Table S9 reports the aggregate paired results for the three score-only base predictors. Gains are computed from unrounded accuracies; displayed accuracies and gains are rounded independently.

Table S9. Aggregate generality across score-only base predictors.

Base predictor	Base accuracy (%)	Base + PCRI (%)	Gain (pp)	95% CI
MPE	64.176	67.190	+3.014	[+2.002,+3.964]
WPE	64.829	68.401	+3.573	[+2.110,+5.345]
CARPRT	67.299	69.984	+2.685	[+1.529,+3.895]

A joint visual summary of base-predictor generality and representative post-score computational cost is provided in Figure S6 in Appendix G.2.

E.2 Unlabeled adaptation-support fraction

This analysis uses ten non-ImageNet datasets with CLIP ViT-B/16 and three fixed seeds (20260841–20260843), yielding 30 setting–seed trials per support fraction. The frozen CARPRT base score

field is held fixed. Samples are uniformly permuted without stratification, and $\mathrm{round}(fN)$ unlabeled samples are supplied only for estimating PCRI's target-centering, ridge map, prototypes, and structural scale. Predictions are evaluated on the full target evaluation pool. Table S10 reports mean paired gains over the same frozen CARPRT field.

Table S10. Sensitivity to unlabeled adaptation-support fraction, reported as mean paired gain over the fixed matched CARPRT score field.

Support fraction	Mean paired gain (pp)
10%	-1.239
25%	+1.827
50%	+2.482
75%	+2.592
100%	+2.685

E.3 PROMPT BUDGET

We evaluate prompt budgets $P \in \{16, 32, 64, 128, 247\}$ over the same ten datasets and the same three seeds (20260841–20260843). Within each seed, prompt subsets are nested prefixes of a fixed random permutation, after which selected prompts are restored to their original order and SHA-256 groups are recomputed. Table S11 reports mean paired gains against the matched CARPRT baseline at each budget.

Table S11. Prompt-budget sensitivity with a matched CARPRT baseline at each budget.

Prompt budget P	Mean paired gain (pp)
16	+2.745
32	+3.049
64	+2.916
128	+2.646
247	+2.685

Absolute-accuracy reconstruction for main Figure 4. Absolute PCRI accuracies are reconstructed as the matched CARPRT accuracy plus the final paired PCRI gain, joined one-to-one by dataset, condition (adaptation-support fraction or prompt budget), and seed. The CARPRT field used for the support-fraction panel is frozen before PCRI support is restricted. Legacy PCRI accuracies from the earlier detail files are not used.

Exact sensitivity profiles. Tables S10 and S11 provide the numerical values underlying the target-support and prompt-budget analyses. Figure S5 presents the same paired gains in a compact visual form. The target-support profile highlights the transition from negative transfer at 10% support to positive gains from 25% onward, whereas the prompt-budget profile remains positive and comparatively stable across substantially different prompt counts.

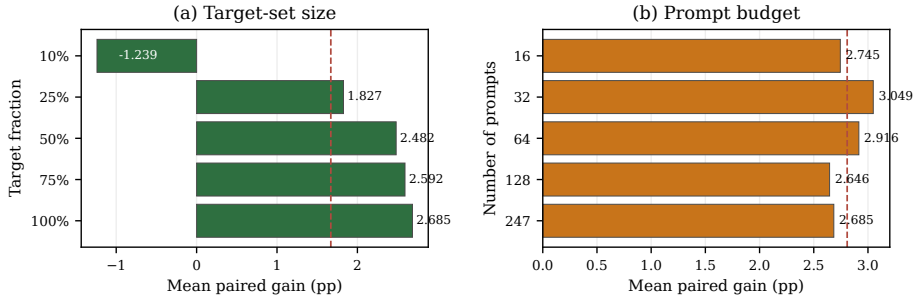


Figure S5. Exact paired-gain sensitivity profiles. (a) Mean gain as unlabeled support varies over a fixed matched CARPRT field. (b) Mean gain across prompt budgets. Dashed lines mark the mean across displayed conditions; values reproduce Tables S10 and S11.

E.4 PROMPT-PARTITION ROBUSTNESS

To assess sensitivity to the prompt partition, G is fixed to 4 and the deterministic SHA-256 grouping is compared with five fixed random partitions over the ten non-ImageNet datasets with CLIP ViT-B/16. Matched CARPRT is 67.299% for every row.

Table S12. Prompt-partition robustness. Performance is similar across deterministic and random partitions; SHA-256 serves as the reproducible default.

Partition	PCRI accuracy (%)	Gain (pp)	W/T/L
SHA-256	**69.984**	**+2.685**	**9/0/1**
Random seed 1	69.980	+2.681	9/0/1
Random seed 2	69.877	+2.578	9/0/1
Random seed 3	69.920	+2.621	10/0/0
Random seed 4	70.030	+2.730	9/0/1
Random seed 5	69.935	+2.635	9/0/1

E.5 DEVELOPMENT-ONLY HYPERPARAMETER SENSITIVITY

All four sweeps use only CIFAR-100 + CLIP ViT-B/16; the development CARPRT reference is 68.590%. No evaluation-dataset label is used to select α, β, G, or λ. The curves show broad high-performing regions rather than a need for per-dataset retuning.

Figure S2. Development-only sensitivity of α, β, G, and λ. Dashed markers indicate the frozen values used in all evaluation settings.

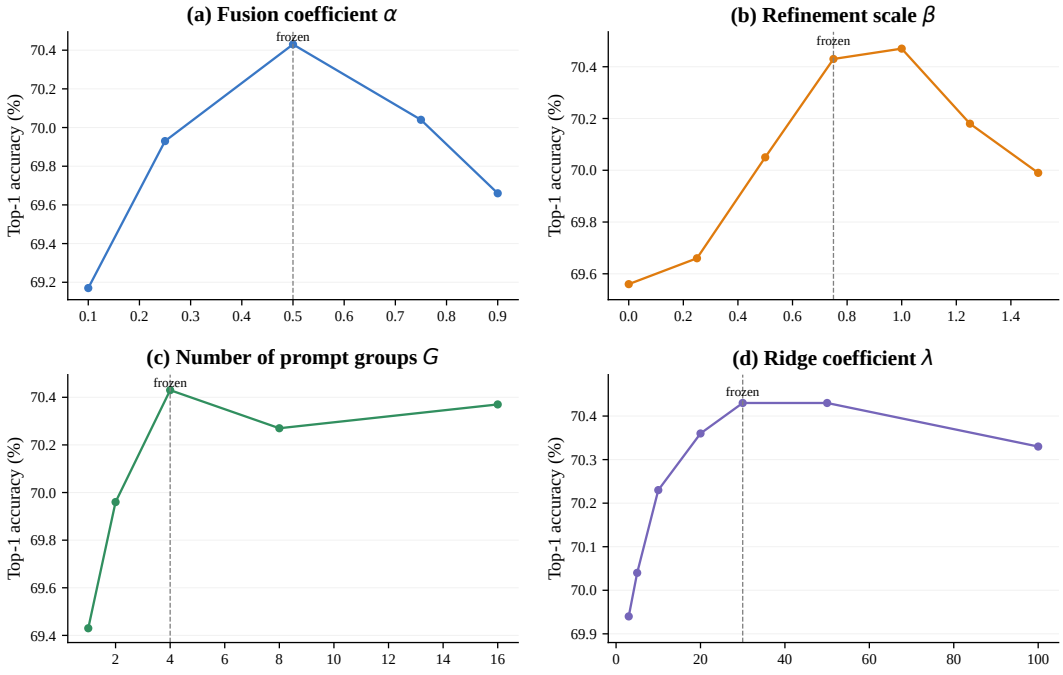


Table S13. Development-set interpretation of the frozen hyperparameters.

Parameter	Frozen value	Development-set interpretation
α	0.50	Best tested development value; 0.25–0.75 all remain positive
β	0.75	Broad high-performance region; $\beta = 1.00$ is +0.040 pp higher on development
G	4	Best tested; $G = 4$–16 forms a stable high-performance plateau
λ	30	Frozen within the 20–50 plateau; $\lambda = 30$ and 50 tie at 70.430%

Development-only prototype aggregation control. Across eight CIFAR-100 + CLIP ViT-B/16 development-only prototype rules, the best alternative reaches 70.520%, only +0.090 pp above the canonical median-support/equal-weight rule (70.430%) and below the pre-specified +0.200 pp advancement threshold; the canonical rule is therefore retained without evaluation-set selection.

F Stress Tests and Failure Boundaries

The former main-text support-composition figure is retained here so that the stronger cross-fitted validation in Figure 5 does not hide PCRI's operating boundary.

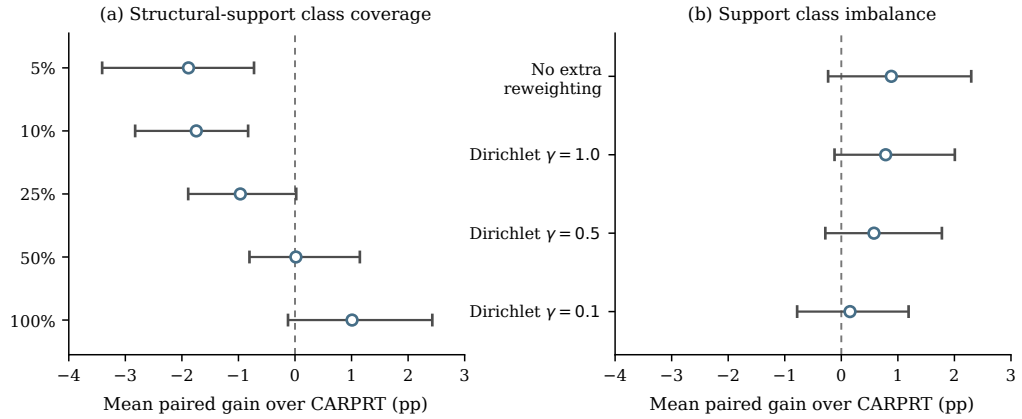


Figure S7: Sensitivity to unlabeled structural-support composition. (a) Low class diversity in the support supplied to PCRI can cause negative transfer. (b) Stronger class imbalance lowers the point estimate, but all intervals cross zero. The frozen base score field is held fixed; construction and uncertainty summaries are given below.

F.1 Structural-support class coverage

Structural-support class diversity is evaluated on Caltech101, Food101, SUN397, and ImageNet with CLIP ViT-B/16. The frozen base score field is computed once from the complete target pool and then held fixed. Five seeds (20260841–20260845) yield 20 trials per coverage fraction. Classes are sampled uniformly without replacement, and samples of selected classes form the support supplied to PCRI for target-state estimation; accuracy is reported on that selected-class support. Ground truth is used only by the evaluation harness to construct this controlled stress test and compute metrics.

Table S14. Structural-support class coverage with the frozen base score field held fixed. Points report mean paired gains and intervals provide 95% uncertainty summaries.

Support class coverage	Mean paired gain (pp)	95% CI
5%	-1.884	[-3.410,-0.725]
10%	-1.745	[-2.826,-0.829]
25%	-0.967	[-1.888,+0.022]
50%	+0.016	[-0.805,+1.147]
100%	+1.009	[-0.124,+2.428]

F.2 Class imbalance

The imbalance analysis uses the same four datasets and backbone, with the frozen base score field held fixed as in Section F.1. Each structural-support trial contains round(0.5N) samples. Symmetric Dirichlet class probabilities are converted to per-sample weights and sampling is without replacement. Five seeds (20260841–20260845) produce 20 trials per condition.

Table S15. Structural-support class-imbalance stress test at 50% support-sample fraction with the frozen base score field held fixed. Points report mean paired gains and intervals provide 95% uncertainty summaries.

Condition	Mean paired gain (pp)	95% CI
No extra reweighting	+0.884	[-0.233,+2.297]
Dirichlet 1.0	+0.784	[-0.120,+2.007]
Dirichlet 0.5	+0.578	[-0.283,+1.778]
Dirichlet 0.1	+0.154	[-0.781,+1.190]

F.3 Coverage-aware fusion under sparse support

The structural-support class-coverage study above shows that PCRI can become unreliable under sparse structural support. To test whether this failure mode can be attenuated using only information visible under the score-only interface, we evaluate a coverage-aware fusion rule as an exploratory robustness extension on the frozen CIFAR-100 development cache. The Score-space Core, prompt residual construction, Core-conditioned ridge residualization, prototype operator, and scale matching are unchanged; only the final fusion strength is adjusted from the pseudo-class coverage observable in B. This extension is evaluated separately from canonical PCRI.

Let the distinct pseudo-class count be the number of classes produced by the frozen base score matrix B over the current unlabeled target batch. We normalize this observable support by the largest number of classes that can be represented by the batch:

$$\rho_{\mathrm{pseudo}} = \frac{K_{\mathrm{pseudo}}}{\min(C, N)}. \tag{F1}$$

$$\alpha_{\mathrm{eff}} = 0.50\, \rho_{\mathrm{pseudo}}^{3}. \tag{F2}$$

The cubic attenuation preserves the original fusion strength when pseudo-class support is broad, while rapidly reducing the structural contribution as the predicted support collapses. Specifically, when the normalized pseudo-class coverage is close to one, the effective fusion coefficient is close to the frozen value 0.50; when the normalized coverage approaches zero, the effective coefficient approaches zero and the prediction smoothly falls back toward the base score matrix B. The resulting fusion is:

$$B_{\mathrm{PCRI,eff}} = (1 - \alpha_{\mathrm{eff}})B + \alpha_{\mathrm{eff}}\widetilde{G}_{\mathrm{struct}}. \tag{F3}$$

Table S16. Development stress-test evaluation of pseudo-class-coverage-aware fusion. Low-coverage statistics summarize controlled sparse-support development batches; the full-set row reports the unchanged CIFAR-100 development accuracy.

Metric	Fixed-α PCRI	Coverage-aware fusion
Low-coverage mean gain	-4.453 pp	+0.017 pp
Worst observed gain	-31.250 pp	-1.250 pp
Negative-transfer rate	57.1%	19.0%
Full CIFAR-100 accuracy	70.430%	70.430%

Coverage-aware attenuation changes the low-coverage mean gain from -4.453 pp to +0.017 pp, reduces the worst observed degradation from -31.250 pp to -1.250 pp, and lowers the negative-transfer rate from 57.1% to 19.0%, while leaving full-set CIFAR-100 development accuracy unchanged at 70.430%. These results show that pseudo-class coverage can be a useful score-only reliability signal under this controlled development stress test. The rule remains an exploratory robustness extension: it is not part of canonical PCRI, is not claimed to be lossless on all benchmark settings, does not provide a universal no-harm guarantee, and should not be interpreted as automatically identifying every source of negative transfer.

F.4 FAILURE-BOUNDARY INTERPRETATION

Proposition F.1 (Observable-equivalent no-harm boundary). Let a base rule h_0 and a score-only adaptation rule h be functions of the same cached score tensor S. If there exists an observable S^* for which $h(S^*) \neq h_0(S^*)$, then, without assumptions restricting $P(Y \mid S)$, h cannot be guaranteed to incur no more sample-wise error than h_0 for every data-generating distribution.

Proof. Choose one data-generating state concentrated on S^* whose label is $h_0(S^*)$, and another observationally identical state whose label is $h(S^*)$. The first favors the base rule and the second favors the adapted rule, although both expose exactly the same S^*. Therefore no decision based only on that observable can distinguish which decision is safe in both states. □

This proposition is deliberately narrow. It excludes an assumption-free, distribution-free, sample-wise no-harm guarantee for any nontrivial score-only decision change; it does not rule out useful average-case guarantees under identifiable structural assumptions.

The failure boundary follows the same target-conditioned representation premise as the main method. PCRI is effective when the unlabeled support used to estimate its target-conditioned state is sufficiently representative, but its structural correction can become unreliable when the population used to estimate centering, pseudo-class support, prototypes, and conditional prompt innovation is extremely sparse or class-partial. The full benchmark contains one negative setting (ViT-B/16 on Caltech101, -0.527 pp), the adaptation-support study is negative at 10% support (-1.239 pp), and the 5% and 10% structural-support class-coverage intervals lie entirely below zero. These observations test the support condition implied by the target-conditioned representation rather than constituting an unrelated post-hoc limitation. More generally, class-equivariant statistics of a cached score tensor need not identify semantic correctness when distinct target states are observationally compatible. This is a strict cached-score observability boundary, not a theorem that rules out every score-only adaptation procedure. Under class imbalance, the mean gain decreases as imbalance strengthens while all intervals cross zero, so that analysis is treated as sensitivity characterization. The coverage-aware rule in Section F.3 is only an exploratory robustness extension motivated by this boundary; canonical PCRI remains unchanged.

G EFFICIENCY AND COMPLEXITY

G.1 ASYMPTOTIC COMPLEXITY

Let $d = (G+1)C$ denote the joint representation dimension and $K \leq C$ the number of valid pseudo-classes. Prompt-level class standardization and prompt-mean removal require $O(NPC)$ operations. For ridge residualization, forming $U^\top U$ costs $O(NC^2)$, forming $U^\top \bar{H}$ and the residual UA^* costs $O(NGC^2)$, and factorizing the $C \times C$ regularized system once and solving for GC right-hand sides costs $O(C^3 + GC^3)$. Prototype construction is $O(Nd)$, while structural scoring is $O(NKd) = O(NK(G+1)C)$; in the worst case $K = C$, this becomes $O(N(G+1)C^2)$. Thus the post-score computation is dominated by the prompt-tensor pass, ridge cross-products/solve, and sample-to-prototype scoring, depending on N, P, C, and K.

Exact sample-space ridge reformulation. The current implementation uses the primal $C \times C$ ridge system. When $N < C$, the same estimator admits the exact identity

$$A^* = (U^\top U + \lambda I_C)^{-1} U^\top \bar{H} = U^\top (UU^\top + \lambda I_N)^{-1} \bar{H},$$

which shifts the matrix factorization from a $C \times C$ system to an $N \times N$ system. This does not reduce the entire ridge stage to $O(N^3)$ because forming the required cross-products and multiplying by $\bar{H}$ still incurs additional matrix-operation costs. When both N and C are large, iterative, low-rank, or active-class approximations are possible extensions; their accuracy–efficiency trade-offs are outside the present evaluation.

G.2 POST-SCORE RUNTIME AND MEMORY

Timing starts after frozen VLM score extraction is complete. Each method/setting uses five warm-up runs followed by 20 timed repetitions with time.perf_counter; torch.cuda.synchronize() is called

after every run. Peak GPU allocation is measured by resetting PyTorch peak-memory statistics before each timed run and reading torch.cuda.max_memory_allocated afterward. Table S17 therefore reports the GPU allocation of the post-score adaptation stage only; it does not measure the storage or disk footprint of the full N x P x C frozen score cache. In particular, the 2.31 GiB ImageNet value should not be interpreted as the memory required to materialize the complete prompt-class score tensor.

Table S17. Representative post-score cost. Score extraction and frozen score-cache storage are outside the reported timing/memory scope; Peak GPU denotes measured allocation during downstream adaptation.

Backbone	Dataset	N	CARPRT ms	Core ms	PCRI ms	PCRI ms/sample	Peak GPU	Extra VLM queries
CLIP ViT-B/16	Caltech101	2,465	0.67	96.05	104.88	0.0425	19.9 MiB	0
CLIP ResNet-50	SUN397	19,850	13.92	737.14	768.91	0.0387	379 MiB	0
DeCLIP ViT-B/32	ImageNet	50,000	74.77	5309.36	5338.57	0.1068	2.31 GiB	0

Generality and post-score cost. Figure S6 provides a visual companion to the aggregate generality results in Appendix E.1 and the representative efficiency measurements above. Panel (a) compares each score-only base predictor with the same predictor augmented by the frozen PCRI configuration. Panel (b) summarizes downstream runtime after the frozen score tensor has already been materialized; it therefore characterizes score-space post-processing rather than end-to-end VLM inference.

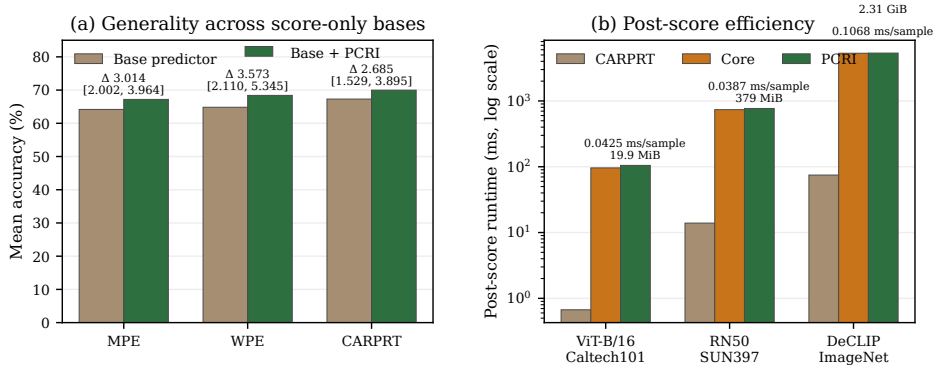


Figure S6. Generality across score-only base predictors and representative post-score efficiency. (a) Mean accuracy of MPE, WPE, and CARPRT before and after applying the same frozen PCRI configuration; annotations report the paired gain and 95% confidence interval. (b) Post-score runtime of CARPRT, the Core pipeline, and full PCRI for three representative dataset–backbone settings. Runtime excludes frozen VLM score extraction, and PCRI issues zero additional VLM queries after score materialization.

G.3 Additional VLM-query accounting

PCRI issues zero additional VLM queries relative to the matched score-only baseline after score extraction. Accordingly, Table S17 measures post-score adaptation cost rather than end-to-end VLM inference latency.

H Reproducibility and Artifact Manifest

H.1 Software and hardware environment

Experiments use Python 3.10.20, PyTorch 2.4.0+cu121, torchvision 0.19, CUDA runtime 12.1, and Windows 10.0.26200. Runtime measurements use an Intel Core i7-13650HX with 23.73 GiB RAM and an NVIDIA GeForce RTX 4060 Laptop GPU (8188 MiB, driver 560.94).

H.2 CHECKPOINT AND PROMPT CHECKSUMS

Table S18. Checkpoint and prompt-bank identifiers and source notes.

Artifact	SHA-256	Source / provenance note
CLIP ViT-B/16	5806E77CD80F8B59890B7E101EABD078D9FB84E6937F9E85E4ECB61988DF416F	Official OpenAI CLIP loader/model source; local bytes pinned by SHA-256.
CLIP RN50	AFEB0E10F9E5A86DA6080E35CF09123ACA3B358A0C3E3B6C78A7B63BC04B6762	Official OpenAI CLIP loader/model source; local bytes pinned by SHA-256.
DeCLIP ViT-B/32	168D4CA6ACA5C454220CBA2CBAC2196C03202A4C096F532ADC9EB2849348BC3E	Official Sense-GVT/DeCLIP upstream model zoo; archived local checkpoint and loader pinned by SHA-256/configuration.
247-template prompt bank	112859FD0FA8CA95603297FDE592DA0A4AFA866F311BEABB2D154087F2AE57E1	Byte-identical to matched CARPRT runtime copy; ordered indices, prompt-string hashes, and deterministic group IDs released with the artifact; template text remains upstream.

H.3 ARTIFACT CONTENTS

Table S19 inventories the companion PCRI code package. Paths are relative to its root (not the Overleaf source directory). It contains the final method implementation and selected result records, not the frozen score caches or every per-trial appendix file.

Table S19. Released PCRI implementation and selected result records.

Artifact entry	Supports
`pcri/operators.py` and `pcri/api.py`	Frozen score-only operators and inference API
`pcri/crossfit.py` and `pcri/sharded.py`	Query-excluded state fitting and large-score execution
`scripts/run_benchmark.py` and `scripts/evaluate.py`	Multi-setting runner and separate post-hoc evaluation
`assets/prompt_bank_sha256_group_mapping.csv`	Ordered indices, prompt-string hashes, and group IDs (no template text)
`results/appendix_33_setting_matched_results.csv`	33 matched CARPRT/PCRI accuracies and paired gains
`results/crossfit/all33_setting_summary.csv`	Per-setting query-excluded gains
`results/paper_controls/core_decomposition_33_detail.csv`	Matched 33-setting Core-decomposition records
`results/paper_controls/semantic_anchor_33_detail.csv`	33-setting semantic-anchor control
`results/paper_controls/predictability_10_detail.csv`	Held-out prompt-predictability diagnostic
`docs/REPRODUCTION.md` and `provenance/source_records.json`	Reproduction scope and source hashes

H.4 REPRODUCTION NOTES AND PROVENANCE

Table S19 lists the released operators, execution entry points, selected result records, and provenance notes. Checkpoint identity is pinned by full SHA-256 hashes and loading configurations. For DeCLIP ViT-B/32, the upstream project and released model zoo are the official Sense-GVT/DeCLIP repository (https://github.com/Sense-GVT/DeCLIP); the exact local checkpoint used in this study is additionally pinned by its archived filename, loader configuration, and SHA-256. This separates upstream provenance from byte-exact local identity and avoids dependence on a machine-local path.

The prompt bank is cryptographically identified and byte-identical to the matched CARPRT runtime copy. The companion artifact includes an ordered index/hash/group manifest but not the template strings, which must be obtained from the identified upstream prompt bank. Controlled subset analyses report their seed ranges and construction rules in Table S3, while the primary dataset-clustered bootstrap is fixed to 100,000 replicates with RNG seed 20260825. Frozen VLM score caches, checkpoint weights, raw images, prompt strings, and some per-trial appendix records are not bundled with the code package. A permanent external mirror or DOI is a release action outside the manuscript itself; no unverified archive identifier is fabricated in this supplement.

H.5 REPRODUCIBILITY SCOPE

The code package pins the score-only implementation, frozen hyperparameters, ordered prompt hash/group mapping, source identifiers, and selected final-paper result records. Full-pool inference and query-excluded cross-fitting are executable when the corresponding external score caches and upstream prompt text are supplied. The cross-fit protocol uses five folds and seeds 20260913–20260915; the reported 33-setting rows are bundled. The exploratory coverage-aware fusion in Appendix F.3 is described and summarized in this supplement but is not part of the released canonical method. Because raw score caches and some appendix trial files are not bundled, the code package alone does not independently rerun every reported experiment.

Table S20 summarizes experimental detail reported in this supplement or the companion package; it does not imply that every underlying trial file is distributed.

Table S20. Supporting records available for each supplementary analysis.

Analysis	Available record
Main matched evaluation	Complete 33-setting result table and bootstrap protocol
Cross-fitted target-state transfer	Complete 33-setting rows; three five-fold allocations; query-exclusion contract; canonical replay checks; dataset-cluster bootstrap
Score-space geometry	Per-dataset values and Figure 1 sampling/plotting protocol reported in Appendix D.3; raw pairwise observations not bundled
Core decomposition and conditioning controls	Matched 33-setting Raw/U_0/Full-Core results; exact control construction; dataset-clustered percentile bootstrap, 20,000 replicates, seed 20260823
Structural-support stress tests	Aggregate tables, seed ranges, construction rules, and bootstrap protocol reported in Appendix F; per-trial files not bundled
Base-predictor generality	Aggregate results and paired uncertainty summaries available
Adaptation-support fraction	Aggregate paired gains, fixed-base protocol, and sampling seeds reported in Appendix E.3
Prompt budget	Aggregate paired gains and sampling protocol reported in Appendix E.3
Development sensitivity, prompt partition, and efficiency	Reported values and experiment protocols available