DENSE SIMILARITY AS A DESIGN CHOICE: CONSISTENCY AND UTILITY IN CLIP READOUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

CLIP's native image–text function does not uniquely determine the internal similarities used by dense readers. We turn this freedom into Learned G, a dense adaptation method that learns task-directed metrics within a family of native-equivalent Q/K implementations. Reciprocal query/key transforms preserve native attention while changing the self-similarities used for dense aggregation. The transforms fold into existing projections without online adapter multiplications. On ViT-B/16, G improves matched SCLIP by 3.49 mIoU points over learned per-head scalars on VOC. Removing isotropic scale retains gains after scalar recalibration; spectrum-preserving rotations reduce performance. Retraining on ViT-L/14 yields gains over per-head scalars of 8.13 points on VOC and 4.89 on a fixed COCO-Object panel. With a trained output head held fixed, learning G upstream adds 1.97 points on VOC, showing that metric adaptation can complement output-side adaptation. Canonicalization gives qualifying readers consistent predictions across equivalent implementations; metric–bias transport retains a selected task across coordinate changes. Dense supervision can thus improve segmentation without changing CLIP's native function.

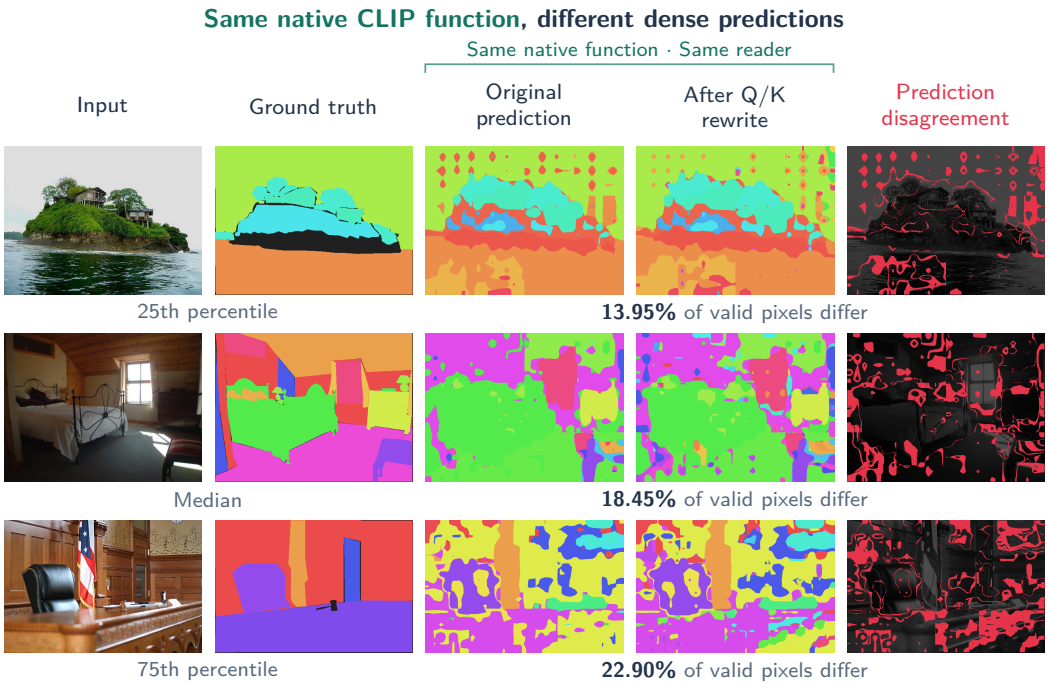

Figure 1: Same native CLIP function and reader, different dense predictions. PEARL PA+KK pre-TLP masks before/after reciprocal Q/K rewriting (ViT-B/16). Examples are nearest the 25th, 50th and 75th disagreement percentiles of a fixed 2,000-image ADE training cohort. Class colors are shared; red marks disagreement, not necessarily errors. Percentages use each image's valid pixels.

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.

1 INTRODUCTION

CLIP's image–text pretraining provides a powerful representation for open-vocabulary recognition (Radford et al., 2021). Dense readers repurpose its internal features for segmentation: SCLIP aggregates through query/key self-correlations, NACLIP combines key similarity with a spatial prior, and PEARL aligns query/key geometry before aggregation (Wang et al., 2024; Hajimiri et al., 2025; Pei et al., 2026). These choices make internal similarity central to dense prediction. Yet the native image–text function does not uniquely determine that similarity: native-equivalent parameters can produce visibly different segmentation masks (Figure 1).

Reciprocal transformations $Q' = QG$ and $K' = KG^{-\top}$ preserve native attention while changing the self-similarities consumed by these readers. This known parameter symmetry (Wang & Wang, 2026b) exposes a useful design space: **parameter choices indistinguishable to the pretrained function can still be selected by downstream supervision**.

This leaves a task-specific question: does a fixed image–text function still admit useful non-scalar structure for dense aggregation? Holding the native function fixed provides a controlled setting for measuring the task value of aggregation-metric choices. We study their contribution beyond similarity rescaling, across backbones, and after the output representation has already been adapted. These questions motivate learning coupled dense metrics within the native-equivalent Q/K family.

Learned G optimizes segmentation while retaining the native image–text function by construction. Bounded per-head transforms fold into the existing projections, which serve both computations. Learned non-scalar structure outperforms per-head scalar calibration; retaining its spectrum but rotating its directions reduces performance. The non-scalar benefit recurs on a second backbone, and learning G upstream of a fixed trained output head provides a further 1.97 mIoU points.

Selecting a useful task metric also raises the question of how to express it across equivalent implementations. We distinguish a source-independent reference readout, obtained through Certified Affine Quotient Canonicalization (CAQC), from retention of a selected task through compatible metric–bias transport. Canonicalization defines a consistent readout, G selects a useful task metric, and transport preserves that selected task.

Our contributions are (i) a native-function-preserving dense adaptation method based on coupled Q/K metrics; (ii) evidence for task-relevant non-scalar structure, its replication on ViT-L/14, and its complementarity with output-side adaptation; and (iii) consistency and compatibility conditions connecting metric selection to canonical reading and fixed-task reuse.

2 RELATED WORK

Dense CLIP adaptation. MaskCLIP reads value-projected features; SCLIP and NACLIP modify self-correlation, while ClearCLIP, ResCLIP, GEM and CLIP Surgery change feature composition or token interaction (Zhou et al., 2022; Wang et al., 2024; Hajimiri et al., 2025; Lan et al., 2024; Yang et al., 2025; Bousselham et al., 2024; Li et al., 2025). PEARL aligns keys to queries before aggregation (Pei et al., 2026). SCLIP also evaluates a learned projection for self-correlation using 64 labeled images, making task-trained similarity a direct precedent. ARM learns a plug-in cross-attention refinement module for existing CLIP segmenters (Liu et al., 2025b). Our specific adaptation space couples query and key metrics reciprocally, so the modified projections preserve native CLIP computation.

Learning within functional equivalence. Reciprocal Q/K symmetries are established properties of attention (Wang & Wang, 2026a;b). Implementation invariance in attribution likewise asks whether equivalent networks yield the same derived output (Sundararajan et al., 2017). SpinQuant learns function-preserving rotations to improve quantized inference (Liu et al., 2025a); we optimize a dense reader sensitive to non-orthogonal Q/K reparameterization. LoRA learns additive low-rank updates and can merge them into pretrained weights without an online branch (Hu et al., 2022). Learned G's distinguishing constraint is native preservation by the rewritten Q/K parameters themselves. Our reader-only LoRA baseline retains the native interface through the original parameter path.

Canonicalization and task reuse. Partial canonicalization leaves residual transformations respected by a predictor (Kaba et al., 2023); related work studies canonical forms, equivariant frames and their continuity (Ma et al., 2024; Dym et al., 2024). Q/K whitening and balancing have been proposed previously (Wang & Wang, 2026b;c). We derive consistency for the analyzed dense readers, including PEARL's finite-step alignment, and characterize when a fixed task metric and effective bias remain compatible with the residual coordinates. These reader-specific results connect known symmetries to dense adaptation and reuse.

3 Native Equivalence Leaves a Dense Metric Choice

The native CLIP function F_θ and dense reader D_θ share parameters but use different aggregation rules. For n input tokens and a head of width d,

$$Q = XW_Q + \mathbf{1}b_Q^\top, \quad K = XW_K + \mathbf{1}b_K^\top, \quad P = \mathrm{softmax}(QK^\top/\sqrt{d}). \tag{1}$$

Matched SCLIP and NACLIP instead aggregate the unchanged values V through

$$\begin{aligned} A_{\mathrm{S}} &= \mathrm{softmax}(sQQ^\top/\sqrt{d}) + \mathrm{softmax}(sKK^\top/\sqrt{d}), \\ A_{\mathrm{N}} &= \mathrm{softmax}(sKK^\top/\sqrt{d} + \Omega), \end{aligned} \tag{2}$$

where $s > 0$ is a similarity multiplier and Ω a fixed spatial prior. For any invertible per-head G, the reciprocal action (Wang & Wang, 2026b)

$$Q' = QG, \qquad K' = KG^{-\top} \qquad \Longrightarrow \qquad Q'K'^\top = QK^\top \tag{3}$$

preserves native attention and the complete image–text function when all remaining computation is fixed. In contrast, with $M = GG^\top \succ 0$,

$$Q'Q'^\top = QMQ^\top, \qquad K'K'^\top = KM^{-1}K^\top. \tag{4}$$

An anisotropic M changes the feature directions contributing to dense similarity. Native-QK and value-only readers provide invariant controls for this intervention.

Dense changes follow the self-similarity path. In a controlled ViT-B/16 audit, a fixed diagonal rewrite changes SCLIP and NACLIP mIoU by -4.888 and -8.780 pp on all 1,449 VOC validation images, and by -2.220 and -3.956 pp on 2,000 ADE images. Native-QK and value-only controls remain unchanged (Appendix C.2.2). On a separate 128-image panel, injecting only the transformed self-metrics reproduces the SCLIP/NACLIP token outputs, patch logits and masks; restoring the original metrics recovers all three with zero recorded error (Appendix C.3). These interventions locate the changing computation in the metrics that G learns.

A separate serialized rewrite with diagonal factors $1/2$ and 2 produces identical normalized embeddings and image–text logits on all 50,000 ImageNet validation images, with no top-1 flips (Deng et al., 2009); Appendix C.2.1 gives the checkpoint and Entmax diagnostics. Algebra establishes native preservation for every input; the finite audits verify the implementations.

A reader is *function-identifiable* if $F_\theta = F_{\theta'}$ implies $D_\theta = D_{\theta'}$. SCLIP and NACLIP violate this property in the controlled audits. Their metric freedom provides a supervised design space within a fixed native function.

4 Learning and Reusing Dense Metrics

4.1 Learned G: Native-Function-Preserving Metric Adaptation

Learned G acts on the Q/K projections of the **final visual Transformer block** (Figure 2). Each of its H attention heads learns a separate transform G_h, shared across images; $H = 12$ for ViT-B/16 and $H = 16$ for ViT-L/14, with head width $d = 64$ in both. Let $\theta(G)$ denote the reciprocal rewrite of the pretrained parameters θ_0. We optimize

$$\min_{G \in \mathcal{G}_r} \ \mathbb{E}_{(x,y)\in\mathcal{D}_{\mathrm{train}}} \mathcal{L}_{\mathrm{seg}}(D_{\theta(G)}(x), y) + R(G), \qquad F_{\theta(G)} \equiv F_{\theta_0}, \tag{5}$$

where $\mathcal{G}_r$ is a bounded per-head family. Reciprocal coupling enforces native equivalence. For the self-Gram readers in Eq. (2), the dense computation depends on G only through $M = GG^\top$, so the SPD representative $G = M^{1/2}$ realizes the same coupled metrics. Low-rank deviations from identity limit the adapted subspace, while bounded log-eigenvalues control conditioning. We parameterize

$$G_h = I + U_h\{\mathrm{diag}(e^{\lambda_h}) - I\}U_h^\top, \quad U_h^\top U_h = I, \quad \lambda_h = \tfrac{1}{2}\log(4)\tanh(a_h). \tag{6}$$

QR supplies $U_h \in \mathbb{R}^{d\times r}$. Each G_h is full rank, has $\mathrm{rank}(G_h - I) \le r$ and condition number at most four, and starts at identity with $a_h = 0$. The $Hr(d+1)$ optimization parameters total 12,480 for G16 on ViT-B/16 and 16,640 on ViT-L/14.

Training, selection and folding. We freeze the pretrained CLIP parameters and text representations and optimize G using differentiable patch logits $z(G)$:

$$\mathcal{L} = \tfrac{1}{2}\mathcal{L}_{\mathrm{CE}}(z(G), y) + \tfrac{1}{2}\mathcal{L}_{\mathrm{softIoU}}(z(G), y) + \frac{10^{-3}}{Hr}\sum_{h,j}\lambda_{hj}^2. \tag{7}$$

The IoU term averages classes present in the patch targets; ignored pixels enter neither task term. After training, development mIoU selects a checkpoint. Strength-calibrated studies also select $\alpha \in \{0.25, 0.5, 0.75, 1\}$, with $\alpha = 0$ retaining the original reader. Strength scales the learned log-eigenvalues:

$$\begin{aligned} G_h(\alpha) &= \exp(\alpha \log G_h) = I + U_h\{\mathrm{diag}(e^{\alpha\lambda_h}) - I\}U_h^\top, \\ G_h(\alpha)^{-\top} &= I + U_h\{\mathrm{diag}(e^{-\alpha\lambda_h}) - I\}U_h^\top. \end{aligned} \tag{8}$$

Training uses $\alpha = 1$; unit-strength studies retain that value at evaluation. For inference, the selected transform (denoted G below) folds into the original projections and biases:

$$W_Q' = W_Q G, \quad b_Q' = G^\top b_Q, \qquad W_K' = W_K G^{-\top}, \quad b_K' = G^{-1} b_K. \tag{9}$$

The same rewritten projections then support native computation and adapted dense reading at their original dimensions, with no online adapter multiplication.

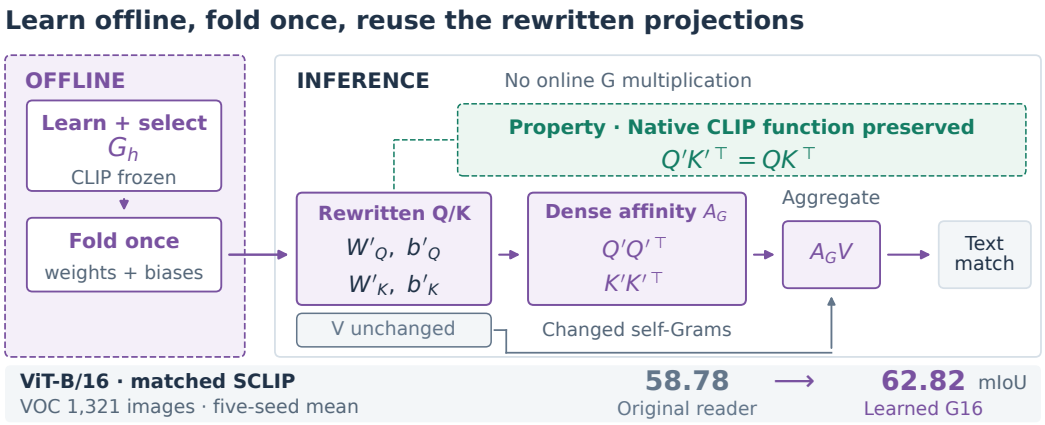

Figure 2: Learning G offline and using its folded projections at inference. Each head in the final visual block learns a transform, selected on the development set and folded once into Q/K weights and biases. Rewritten self-Grams drive dense aggregation of unchanged V; the dashed green annotation states preservation of the native CLIP function. No online G multiplication is required. Scores compare the original reader and G16 (rank budget $r = 16$) from the strength-calibrated projection study in Table 3: five-seed means on 1,321 VOC images, ViT-B/16, matched SCLIP. Checkpoint and strength are selected on dev64 within a six-epoch search; see the reader-only projection protocol in Appendix A.5.1.

Numerical check on learned transforms. For five learned ViT-L/14 states and 12 images, the folded image encoder has a maximum relative Frobenius error of 3.84×10^{-7} in image encodings. Explicit and independently folded dense paths differ by at most 1.55×10^{-6} in class scores, with one decision-boundary pixel difference across the 60 image–state comparisons (Appendix A.4.2).

4.2 Canonical Reading and Its Relation to Learned G

Learning a metric changes the dense task rule. Expressing a selected rule in different Q/K coordinates instead requires its metric and effective bias to transform with those coordinates. We distinguish a source-independent canonical reference from compatible reuse of a selected task.

To obtain a consistent reader across equivalent implementations, we also consider affine key translations. Adding $\mathbf{1}c^\top$ to K changes native logits only by row constants, preserving attention and the centered interaction $QK^\top H_n$, where $H_n = I - \mathbf{1}\mathbf{1}^\top/n$. CAQC removes key bias and then balances the weight Grams by a reciprocal rewrite. For full-column-rank W_Q, W_K, let $S_Q = W_Q^\top W_Q$, $S_K = W_K^\top W_K$ and

$$T = S_Q^{-1/2}(S_Q^{1/2} S_K S_Q^{1/2})^{1/2} S_Q^{-1/2}, \qquad B = T^{1/2}. \tag{10}$$

Rewriting by B, including query bias, balances the Grams since $TS_QT = S_K$. Prior work considers Q/K Gram balancing (Wang & Wang, 2026b;c). We derive the positive-definite map above in Appendix B.2 and use its residual symmetry to characterize complete readers.

Proposition 1 (Readout consistency after affine canonicalization)**.** *Let C be exact CAQC. For θ, θ' in the same per-head affine Q/K orbit, with full-column-rank query/key weights and all other parameters fixed, their canonicalized Q/K pairs differ only by shared orthogonal actions. If the complete dense reader is invariant to these actions on the admitted domain, then $D_{C(\theta)}(x) = D_{C(\theta')}(x)$ for every admitted input.*

The self-Grams in Eq. (2) meet this condition. PEARL's evaluated finite Newton–Schulz alignment also does: each iteration transforms by orthogonal conjugacy, including at singular cross-covariances. Appendix B.2 gives the proofs and reader conditions.

The proposition clarifies the relationship to learning G. For a qualifying reader, $D_{C(\theta(G))} = D_{C(\theta_0)}$: canonicalization removes the orbit-dependent difference selected by G. Canonical reading and task-directed metric selection therefore answer distinct design questions. Retaining an already selected task across new coordinates instead requires carrying its metric and effective bias.

4.3 Compatible Reuse of a Selected Task

Let canonical row features satisfy $Z' = ZO$, $O^\top O = I$. An affine task branch with metric $M \succ 0$ and effective bias β has affinity

$$A_{M,\beta}(Z) = \mathrm{softmax}\left(\frac{ZMZ^\top + \mathbf{1}\beta^\top Z^\top}{\sqrt{d}} + \Omega\right). \tag{11}$$

This follows from the self-Gram of $ZL + \mathbf{1}b^\top$ after dropping row constants, with $M = LL^\top$ and $\beta = Lb$.

Proposition 2 (Compatibility of a fixed affine task affinity)**.** *For fixed finite Ω and $\mathrm{rank}(H_nZ) = d$,*

$$A_{M,\beta}(ZO) = A_{M,\beta}(Z) \quad \Longleftrightarrow \quad OMO^\top = M, \quad O\beta = \beta. \tag{12}$$

At any feature rank, $M' = O^\top MO$ and $\beta' = O^\top \beta$ give $A_{M',\beta'}(ZO) = A_{M,\beta}(Z)$.

Our Certified Task-Aware Gauge Compiler (CTGC) implements fixed-task reuse through a canonical reference and an admissible pivot–polar frame. It stores the map to the selected Q/K parameters and target key bias, then folds them into an equivalent source. On the admitted rank and conditioning domain, this recovers the selected affine Q/K state and hence its complete predictions. Appendix B.3 gives the compatibility proof and lower-rank condition; Appendix B.4 gives the construction. Section 5.6 checks prediction retention.

5 Experiments

5.1 Experimental Setup

Models, readers and data. We use released OpenAI CLIP ViT-B/16 and ViT-L/14. Matched SCLIP/NACLIP share preprocessing, text embeddings and decoding, with text scale 40 and background threshold 0.1; separate reuse studies retain complete ResCLIP and PEARL pipelines. Learned

G and adaptation baselines use dense supervision. VOC uses 256 training and 64 development images from the official VOC2012 train split, and 1,321 evaluation images from its validation split (Everingham et al., 2010).

For cross-dataset evaluation, we fix 1,000 COCO train2017 images (Lin et al., 2014) with COCO-Stuff annotations (Caesar et al., 2018), converted to **COCO-Object**: 80 object categories plus background, following the ResCLIP conversion (Yang et al., 2025). All 81 classes enter mIoU. Each backbone's frozen text encoder builds the target text representations from the fixed class names, aliases and prompt templates. Adaptation and selection use VOC only; no target parameters are fitted on COCO-Object. Earlier external replay uses separate COCO-Object and ADE20K panels (Zhou et al., 2017).

Evaluation and uncertainty. We pool image confusion matrices before averaging class IoUs, then average mIoU across seeds. Scores are percentages; differences are percentage points (pp), computed before rounding. Paired-seed Student-t intervals describe training variability on a fixed panel; paired-image bootstrap intervals condition on the selected models. Multiple contrasts use Bonferroni adjustment within each stated family, targeting nominal 95% family-wise coverage; captions give the per-contrast levels. Appendix A details training and selection budgets for each comparison.

The VOC evaluation images were used in earlier experiments. The COCO-Object panel was unused before the ViT-B/16 study and is reused for ViT-L/14. Models are selected on development data before each reported evaluation; Appendix D details data reuse and model selection.

5.2 Non-Scalar Structure and Learned Directions

Metric learning adds more than calibration. After development selection of a global similarity multiplier, G8 improves matched SCLIP and NACLIP by 3.081 and 3.601 pp, respectively (Table 1). A stronger SCLIP baseline learns a separate positive multiplier for each head's QQ and KK branch, totaling 24 scalars. G16 exceeds this baseline by 3.488 pp under matched training and selection budgets.

Table 1: Metric controls on ViT-B/16 and the 1,321-image VOC evaluation set. Global-scalar rows use three seeds and paired-image intervals at 98.75% per contrast (family of four). The head-scalar row uses five paired seeds and a 95% interval; shape rows use five paired seeds and 99.1667% intervals per contrast (family of six). Each block retains its own training protocol. Bold marks the higher mean within each top-block comparison and the two focal shape differences below.

Comparison	Reference	Adapted	Difference, pp [interval]
SCLIP: global scalar → G8	58.782	**61.863**	+3.081 [2.466, 3.698]
NACLIP: global scalar → G8	59.648	**63.249**	+3.601 [3.220, 4.008]
SCLIP: head scalars → G16	59.013	**62.502**	+3.488 [3.065, 3.911]
Expanded scalars S → original G16	59.003	62.502	+3.498 [2.791, 4.206]
Expanded scalars S → learned shape L	59.003	62.329	**+3.325** [2.454, 4.196]
Rotated shape R → learned shape L	58.870	62.329	**+3.459** [2.418, 4.500]

Shape and orientation both carry task value. We remove each selected G's mean log-scale, fixing its shape as $B = \exp(\log G - \mathrm{tr}(\log G)I/d)$. Learned-shape L fixes B; the scalar-only control S fixes $B = I$; rotated-shape R fixes $OBO^\top$ with one prespecified orthogonal draw per seed and head. These diagnostic arms independently refit 24 reader-side QQ/KK scalars with the same recalibration budget and warm start; L and R inherit their shapes from preceding G training (Appendix A.3.2). L exceeds S by 3.325 pp. Rotating the shape while preserving its spectrum lowers performance by 3.459 pp relative to L. Thus utility depends on the relationship between the metric and feature directions, beyond overall scale or eigenvalue magnitudes alone (Figure 3).

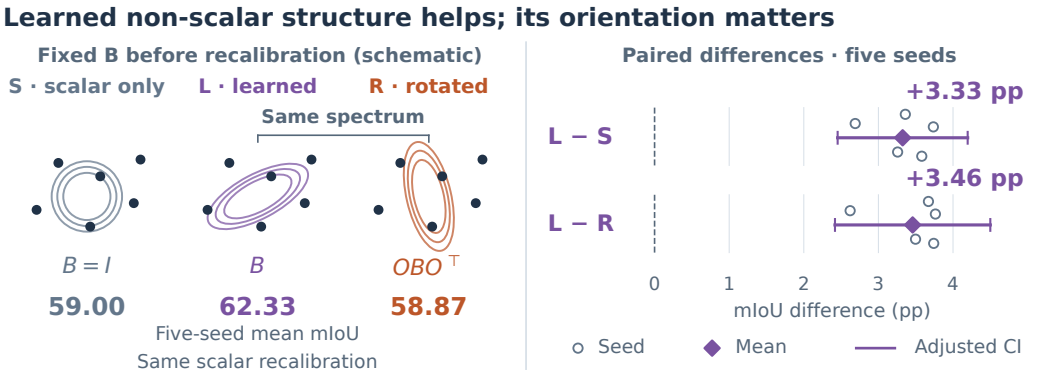

Figure 3: Structure and direction in matched SCLIP, ViT-B/16, on 1,321 VOC images. Left: schematic features with scalar-only S, learned shape L and rotated shape R; L/R have the same fixed-shape spectrum before recalibration. Scores are five-seed mean mIoU. All arms receive the same scalar recalibration budget; L/R inherit their shapes from preceding G training. Right: circles are five paired-seed differences; diamonds and bars give means and 99.1667% per-contrast Student-t intervals (six-comparison Bonferroni family, nominal 95% family-wise coverage).

5.3 Replication on ViT-L/14 and Frozen Reuse

We fit G16 anew on ViT-L/14, alongside 32 branch-wise head scalars, using five seeds and matched six-epoch training and selection budgets. On VOC, G exceeds head scalars by 8.133 pp [7.837, 8.429]; on COCO-Object, without target fitting, the gain is 4.891 pp [4.163, 5.620]. Gains over the original reader are 15.553 and 11.976 pp, respectively. For all four contrasts, the adjusted lower bounds exceed 0.5 pp; each interval has 98.75% confidence under the four-comparison Bonferroni correction. Table 2 gives absolute scores, including reader-only LoRA4; Appendix A.4.1 gives the training protocol and per-seed results.

Table 2: Matched SCLIP on ViT-L/14, mean mIoU over five seeds. All trained models are selected on the 64-image VOC development set. G minus reader-only LoRA4 is -6.500 $[-7.581, -5.419]$ pp on VOC and $+0.592$ $[-0.892, 2.076]$ on COCO-Object. These paired-seed intervals use 97.5% confidence per contrast (family of two). Bold indicates the highest mean within each matched comparison.

Evaluation	Original reader	Head scalars	Learned G16	LoRA4
VOC (1,321 images)	42.417	49.837	57.970	**64.470**
COCO-Object (1,000)	19.455	26.540	**31.431**	30.839

Frozen metrics also support reuse across images and readers. On ViT-B/16, VOC-selected G16 exceeds the fixed head-scalar baseline on COCO-Object by 1.668 pp [0.979, 2.226], with no target fitting. In a separate cross-reader study, fixed NACLIP-trained G16 plus target-reader scalar selection raises complete ResCLIP from 61.066 to 63.521 mIoU and complete PEARL from 64.026 to 64.874. Their gains are 2.455 pp [2.023, 2.868] and 0.848 pp [-0.009, 1.639], respectively; the PEARL increment remains unresolved. These paired-image intervals use 99.1667% confidence per contrast in their respective six-comparison families. Appendix A.7 gives these protocols and the smaller G8 gains on COCO-Object and ADE without target fitting.

5.4 Comparison with Other Adaptation Choices

A text-compatible output head transforms aggregated features before normalization and text matching. We use a rank-r residual $H(Y) = Y + s(YA^\top)B^\top$, with $A \in \mathbb{R}^{r\times 512}$, $B \in \mathbb{R}^{512\times r}$ and $r = 12$ on ViT-B/16.

Reader-only full-projection Q/K LoRA4 and a standalone Hout12 both obtain higher VOC segmentation accuracy than G under strength-calibrated selection (Table 3). Each comparison retains its separately trained G baseline (Appendix D.1). G preserves native attention through the same rewritten Q/K projections used for dense adaptation.

Table 3: Strength-calibrated adaptation on ViT-B/16 and the 1,321-image VOC evaluation set. Each comparison uses five seeds and candidate selection within six epochs. Intervals are paired-seed intervals at 98.3333% per contrast (three comparisons per study). The original reader scores 58.782. G baselines are separately trained; the head here is selected independently of Table 4. Bold indicates the highest mean within each matched comparison.

Comparator	G16	Comparator	G minus comparator, pp [interval]
Reader-only LoRA4	62.823	**64.544**	-1.721 $[-2.860, -0.582]$
Standalone Hout12	62.770	**66.664**	-3.894 $[-5.766, -2.022]$

5.5 Complementarity with Output-Side Adaptation

Output-side adaptation is already effective in the preceding comparison. To test whether pre-aggregation metrics offer additional value, we use a separate fixed-head study. We freeze five selected Hout12 heads and learn G upstream. The result improves from 66.476 to 68.447 mIoU, a 1.971 pp gain (Table 4). G also improves over equal additional-budget controls that continue training the head or add a second output residual. These comparisons support a complementary role for pre-aggregation metric learning (Appendix A.6).

Table 4: Additional adaptation with selected Hout12 heads on the 1,321-image VOC evaluation set, ViT-B/16. Trained arms receive the same four-epoch search budget. Differences compare Head+G with each reference over five paired seeds. The fixed-head comparison uses a 95% interval; the two additional-budget comparisons use 97.5% per-contrast intervals. The two additional-budget intervals exclude zero, while gains above 0.5 pp remain unresolved. Bold highlights the primary fixed-head comparison.

Method	mIoU	Head+G minus reference, pp [interval]
Fixed head	66.476	**+1.971** [0.979, 2.962]
Continue head training	66.676	+1.771 [0.429, 3.114]
Add output residual	67.281	+1.167 [0.358, 1.975]
Fixed head + Learned G	68.447	—

Jointly fitting G and a head from identity gives a different result: G16+Hout12 reaches 67.118, versus 66.476 for Hout12 and 65.799 for Hout24, with both paired increment intervals crossing zero (Appendix C.11). The established complementarity therefore concerns the fixed-head procedure.

5.6 Consistent Reading and Retention of a Selected Task

CAQC realizes readout consistency. On 1,449 VOC images, combined affine rewrites change 7.675% of official NACLIP predictions and 9.099% of official ResCLIP predictions. CAQC reduces disagreement against the canonical reference to at most 0.0001%. Its canonical identity changes mIoU by -0.112 and $+0.140$ pp, respectively. The earlier Weight-Gram construction loses 0.790 pp $[-0.941, -0.628]$ in matched PEARL despite consistent reading. These results separate canonical agreement from metric utility (Appendix C.4).

Transport retains the chosen dense task. We freeze one NACLIP G8 endpoint with 62.0467 mIoU on 128 VOC images, versus the original reader's 58.7915. Across five affine sources, CAQC-only predictions agree with their canonical reference to within four pixels. Nevertheless, naively reusing the selected task changes 3.6140% of valid pixels and loses 2.6517 pp on the fixed orthogonal source. Compatible transport through CTGC retains the selected endpoint across all five sources, with at most two pixels differing from the selected target predictions and maximum mIoU drift 1.594×10^{-5} pp;

full target Q/K copying is exact. Other sources yield at most 0.086% naive disagreement. The replay verifies prediction preservation under the tested equivalent reparameterizations (Appendix C.12).

6 Discussion and Limitations

Native-function preservation leaves room for useful dense adaptation because the two computations use Q/K differently. G constrains the native cross-interaction while segmentation supervision chooses the self-similarities used for aggregation. The preservation guarantee and segmentation accuracy describe separate properties of the resulting model.

The fixed-head result also clarifies where useful adaptation remains. An output head transforms features that have already been aggregated; G changes the token weights before aggregation. Improving a fixed, trained head by 1.97 points shows that the earlier aggregation step still offers useful adjustments. The comparisons with continued head training and an additional output residual further support studying these two adaptation locations together.

This task-directed metric choice also differs from changing its coordinate expression. Canonical reading defines a common reference; transporting the selected metric and effective bias retains the task rule across equivalent implementations.

The evidence covers two backbones, the stated readers and bounded training budgets. The guarantees apply to the analyzed Q/K family and reader/rank conditions. Direction interventions change several attention properties together, so they establish task value in orientation without isolating a unique semantic mechanism. Fixed-task replay evaluates prediction retention under controlled equivalent reparameterizations.

7 Conclusion

Learned G adapts CLIP's dense readout through Q/K transformations that preserve its native image–text function. Non-scalar controls, results on two backbones and gains with fixed output heads show that downstream supervision can use this parameter freedom to improve segmentation. Within the analyzed Q/K family, canonicalization gives qualifying readers consistent predictions, while metric–bias transport retains a selected task across equivalent implementations. Dense similarity therefore remains a task-directed design choice even when the native image–text function is fixed.

AI use statement

Generative AI tools assisted hypothesis critique, experimental design, method implementation, literature lookup, result interpretation, mathematical formulations, derivations and proof checks, manuscript drafting, and figure preparation. The authors are responsible for the final text, claims, and materials.

Reproducibility statement

The supplementary material specifies parameterizations, data and selection boundaries, and statistical units. The accompanying anonymous archive `reproduction.zip` provides configurations, selected-state records, sufficient statistics and CPU recomputation scripts for Tables 1–4, together with the source data and reconstruction instructions for Figures 1–3. Its `paper_index.csv` maps each result to files and commands.

References

Sule Bai, Yong Liu, Yifei Han, Haoji Zhang, Yansong Tang, Jie Zhou, and Jiwen Lu. Self-calibrated CLIP for training-free open-vocabulary segmentation. *IEEE Transactions on Image Processing*, 2025. URL https://arxiv.org/abs/2411.15869.

Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3828–3837, 2024.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1209–1218, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Caesar_COCO-Stuff_Thing_and_CVPR_2018_paper.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Nadav Dym, Hannah Lawrence, and Jonathan W. Siegel. Equivariant frames and the impossibility of continuous canonicalization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 12228–12267, 2024. URL https://proceedings.mlr.press/v235/dym24a.html.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

David Groisser, Sungkyu Jung, and Armin Schwartzman. Uniqueness questions in a scaling-rotation geometry on the space of symmetric positive-definite matrices, 2017. URL https://arxiv.org/abs/1702.03237v2. Version 2.

Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5061–5071, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 15546–15566, 2023. URL https://proceedings.mlr.press/v202/kaba23a.html.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ClearCLIP: Decomposing CLIP representations for dense vision-language inference. In *European Conference on Computer Vision*, pp. 143–160, 2024.

Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 162:111409, 2025. doi: 10.1016/j.patcog.2025.111409.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context, 2014. URL https://arxiv.org/abs/1405.0312.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. SpinQuant: LLM quantization with learned rotations. In *International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e5b1c0d4866f72393c522c8a00eed4eb-Paper-Conference.pdf.

Ziquan Liu, Zhewei Zhu, and Xuyang Shi. Arm: A learnable, plug-and-play module for clip-based open-vocabulary semantic segmentation, 2025b. URL https://arxiv.org/abs/2512.24224.

George Ma, Yifei Wang, Derek Lim, Stefanie Jegelka, and Yisen Wang. A canonicalization perspective on invariant and equivariant learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/702b67152ec4435795f681865b67999c-Abstract-Conference.html.

Gensheng Pei, Xiruo Jiang, Xinhao Cai, Tao Chen, Yazhou Yao, and Byeungwoo Jeon. PEARL: Geometry aligns semantics for training-free open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17927–17937, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3319–3328, 2017.

Feng Wang, Jieru Mei, and Alan L. Yuille. SCLIP: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, pp. 315–332, 2024.

Hong Wang and Kelly Wang. Gauge fiber bundle geometry of transformers. In *Proceedings of the 4th (2025) and 3rd (2024) NeurIPS Workshops on Symmetry and Geometry in Neural Representations*, volume 282 of *Proceedings of Machine Learning Research*, pp. 664–684, 2026a. URL https://proceedings.mlr.press/v282/wang26b.html.

Hong Wang and Kelly Wang. Complete characterization of gauge symmetries in transformer architectures. In *Proceedings of the 4th (2025) and 3rd (2024) NeurIPS Workshops on Symmetry and Geometry in Neural Representations*, volume 282 of *Proceedings of Machine Learning Research*, pp. 622–663, 2026b. URL https://proceedings.mlr.press/v282/wang26a.html.

Hong Wang and Kelly Wang. GaugeKV: Composable exact KV cache compression. OpenReview preprint, 2026c. URL https://openreview.net/forum?id=rSxYPLzyBu. Public ICLR 2026 submission, not accepted proceedings.

Yuhang Yang, Jinhong Deng, Wen Li, and Lixin Duan. ResCLIP: Residual attention for training-free dense vision-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29968–29978, 2025.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 633–641, 2017.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *European Conference on Computer Vision*, pp. 696–712, 2022.

SUPPLEMENTARY MATERIAL

The supplement follows the main argument. Appendix A gives the method and principal experimental protocols; Appendix B gives proofs and coordinate-handling conditions; Appendix C collects further analyses and alternatives; Appendix D provides implementation, statistical and reproduction details.

Main-text result or claim	Reproduction or proof
Table 1, Figure 3: scalar, shape and direction	Sections A.3 and A.3.2
Table 2: second backbone and COCO-Object	Sections A.4 and A.1.1
Table 3: alternative adapters	Sections A.5 and D.1
Table 4: fixed-head complementarity	Section A.6
Section 4.1: folded learned transforms	Section A.4.2
Propositions 1 and 2	Sections B.2 and B.3
Section 3: readout and path interventions	Sections C.2.2 and C.3
Section 5.6: consistency and task retention	Sections C.4 and C.12

A LEARNED G AND MAIN-EXPERIMENT REPRODUCTION

This section follows the main experiments: shared inputs, metric controls, the second backbone, alternative adapters, fixed output heads and frozen reuse. Each comparison retains its own trained states, selection budget and uncertainty unit. The guide above links each main table and proposition to its details.

A.1 DATA, READERS AND SCORING

The supervised adaptation studies use 256 training and 64 development images from the official VOC2012 training split, and 1,321 evaluation images from its validation split. We denote these fixed sets by train256, dev64 and eval1321. All development decisions precede the corresponding evaluation; earlier uses of the evaluation images and post-hoc comparisons are described in Section D. Unless stated otherwise, class confusion matrices are pooled before computing mIoU, and seed scores are averaged rather than ensembling predictions.

The matched reader preprocesses an image with resize factor $\min(2048/w, 336/h)$, rounds resized dimensions as in the pinned implementation, and uses 224-pixel crops with stride 112. Patch sizes are 16 pixels for ViT-B/16 and 14 pixels for ViT-L/14. The fixed text pipeline has 56 aliases for VOC foreground labels, text scale 40, and a background threshold of 0.1. Attention includes CLS and self sources unless the source-scope diagnostic explicitly holds their contributions fixed. The matched utility studies do not use TLP, PA, PAMR, SAM, or DINO.

A.1.1 FIXED COCO-OBJECT PANEL

The 1,000-image COCO-Object panel uses COCO train2017 images and COCO-Stuff164k pixel annotations (Caesar et al., 2018), converted with the frozen ResCLIP object-label mapping. It was locked before its first ViT-B/16 evaluation. Candidate JPEG/PNG pairs were ordered by the frozen hash rule; previously used image IDs and byte hashes were excluded using project records. The selected images and masks were then hash-locked. For scoring, the frozen 81-class conversion maps source PNG label 255 to background; only later crop padding is ignored. All 81 classes remain in the mIoU denominator, including a class with no ground-truth support in this panel. The five fixed G/head-scalar pairs are scored separately by pooling image confusion matrices, then their mIoUs are averaged. No predictions are ensembled across seeds.

A.2 METRIC PARAMETERIZATION AND SELECTION

Learned G uses the final visual block, one transform per head, and the parameterization, loss and folding rules in Eqs. (6) to (9). The pretrained visual parameters and text representations are frozen. G16 has 12,480 optimization parameters on ViT-B/16 and 16,640 on ViT-L/14. A seeded basis is orthonormalized by QR, and zero log-eigenvalues initialize the transform at identity.

Training uses unit strength. In calibrated comparisons, selection considers the stated checkpoint epochs and four strengths $\{0.25, 0.5, 0.75, 1\}$, plus the unchanged reader. G strength multiplies the bounded log-eigenvalues; the inverse uses their negatives. Unit-strength comparisons and scalar-only recalibration retain their separately specified candidate sets. The following subsections give the learning rates, epochs and candidate budgets for each reported comparison.

A.3 Scalar, Shape and Direction Controls

A.3.1 Global and Head-Wise Scalar Baselines

Global scalar selection uses $\{1/8, 1/4, 1/2, 1, 2, 4, 8\}$ on dev64, followed by three-epoch G8 fitting at the selected scalar with two learning rates and three seeds. The selected multipliers are one for SCLIP and 0.5 for NACLIP. The directly learned head-wise baseline has 24 independent QQ/KK log-scalars, initialized at zero and projected to $[-\log 16, \log 16]$ after each update. Its four rates are $\{0.002, 0.01, 0.05, 0.2\}$; the reused G16 states used $\{0.0005, 0.002, 0.005, 0.02\}$. Each method/seed has three epochs, 192 updates per trajectory and 13 candidates including identity. Both share the task loss and keep the backbone frozen; G retains its own spectral regularizer.

In the expanded scalar/shape study, each arm starts from the selected head-scalar warm start, resets optimizer state, and uses rates $\{0.002, 0.02, 0.2, 2\}$ over eight epochs. The 18 candidates comprise identity, the original scalar reference, and four rates at epochs 2, 4, 6 and 8. This gives the scalar control a larger calibration opportunity while holding the learned or rotated non-scalar shapes fixed. The fixed shapes inherit the preceding G training; their acquisition is not included in the new scalar-fitting budget.

A.3.2 Fixed Non-Scalar Shapes and Spectrum-Matched Directions

We fix five previously selected G endpoints and compare their non-scalar shapes on the historical VOC train256/dev64/eval1321 split. For each head, write $H = \log G$, remove its isotropic component as $S = H - (\mathrm{tr}(H)/64)I$, and fix the learned shape $L = \exp(S)$. The randomized control uses $R = OLO^\top$ with one prespecified orthogonal draw per seed and head, preserving the spectrum. The scalar-only, learned-shape and randomized-shape arms each fit 24 scalar parameters from the same frozen scalar warm start, using four learning rates and eight epochs. Full dev64 evaluations at epochs 2, 4, 6 and 8 select the endpoints before eval1321 is accessed. Each seed pools the evaluation confusion matrices before computing mIoU; paired-seed t intervals use four degrees of freedom and retain the prespecified family-of-six correction (Table 5).

For head h, let B_h be the fixed identity, learned or rotated shape and set $\widetilde{Q}_h = Q_h B_h$, $\widetilde{K}_h = K_h B_h^{-\top}$. The reader uses

$$A_h = \mathrm{softmax}\Big(e^{\theta_{q,h}} \widetilde{Q}_h \widetilde{Q}_h^\top / \sqrt{d}\Big) + \mathrm{softmax}\Big(e^{\theta_{k,h}} \widetilde{K}_h \widetilde{K}_h^\top / \sqrt{d}\Big), \qquad d = 64.$$

The 24 fitted log-scalars act only in the dense reader, independently for the QQ and KK branches; they are not folded into the native Q/K projections. The fixed shape itself has a reciprocal implementation. These diagnostic endpoints therefore differ from the single-transform G16 method evaluated in the direct head-scalar comparison.

Table 5: Fixed-shape and scalar controls. Five paired seeds; 99.1667% t intervals retain the family-of-six correction. Differences are mIoU points.

Contrast	Mean difference	Interval
G16 - S	+3.498	[2.791, 4.206]
L - S	+3.325	[2.454, 4.196]
L - R	+3.459	[2.418, 4.500]

A.4 SECOND-BACKBONE CONFIRMATION

A.4.1 VIT-L/14 TRAINING AND SELECTION

We use OpenAI ViT-L/14 with QuickGELU at 224-pixel crop size: width 1,024, 16 heads of width 64, a 16×16 patch grid, and 768-dimensional text-compatible outputs. Preprocessing retains the matched reader's sliding-window rule, text scale 40 and background threshold 0.1. Text embeddings are generated with this backbone's text encoder for the frozen VOC and COCO vocabularies. Training labels use nearest-neighbor patch-target mapping. No canonicalization is applied before learning G.

The arms are G16 (16,640 parameters), T32 (32 independent QQ/KK head scalars) and P4 (18,432 full-projection LoRA weight/bias parameters). T32 optimizes log-scalars projected to $[-\log 256, \log 256]$ after each step. G starts at identity with a seeded Gaussian basis followed by QR; T starts at zero log-scale; P4 starts with zero output factors and bias increments. Each learning rate starts from the same method/seed initialization. The five seeds are 20260821, 20260831, 20260841, 20260851 and 20260861.

G, T32 and P4 use rates $\{0.002, 0.005, 0.02, 0.08\}$, $\{0.002, 0.02, 0.2, 2\}$ and $\{0.0005, 0.002, 0.005, 0.02\}$, respectively. All 60 trajectories run six epochs over train256, with four-image gradient accumulation and 64 updates per epoch (23,040 updates total). AdamW uses $(0.9, 0.999)$, $\epsilon = 10^{-8}$, zero weight decay, constant learning rate and gradient-norm clipping at five. Computation is FP32 with autocast and TF32 disabled; the backbone remains in evaluation mode. Per-seed epoch order is shared across arms and rates, using PCG64(seed$+40000+$zero-based epoch). All arms share the CE/soft-IoU task loss; G alone adds the spectral regularizer in Eq. (7).

Development observations at epochs 1, 2, 3 and 6 use strengths $\{0.25, 0.5, 0.75, 1\}$ plus the original reader, totaling 65 candidates per arm and seed. Strength scales bounded G log-eigenvalues, T log-scalars, or P4 additive weight/bias updates. Pooled dev64 mIoU determines selection; ties within 10^{-9} pp prefer identity, earlier epoch, smaller learning rate, strength nearer one, then stable candidate ID. All 15 selected states are fixed before VOC eval1321 and COCO1000 evaluation. All five G16 endpoints select learning rate 0.08, and all five T32 endpoints select 0.2.

Table 6: Prespecified ViT-L/14 contrasts. Five paired seeds on each fixed panel; the four primary comparisons use 98.75% per-contrast Student-t intervals and the two secondary comparisons use 97.5% per-contrast intervals. Each family targets nominal 95% family-wise coverage by Bonferroni adjustment; all intervals use four degrees of freedom.

Family	Contrast	Difference, pp	Interval
Primary	G−T32, VOC	+8.133	[7.837, 8.429]
	G−T32, COCO	+4.891	[4.163, 5.620]
	G−N, VOC	+15.553	[15.264, 15.843]
	G−N, COCO	+11.976	[11.445, 12.507]
Secondary	G−P4, VOC	−6.500	[−7.581, −5.419]
	G−P4, COCO	+0.592	[−0.892, 2.076]

Every primary paired difference is positive, and all four lower bounds exceed the prespecified 0.5 pp threshold. The secondary comparison favors P4 on VOC and is unresolved on COCO. Per-seed absolute scores are in Table 7; these repetitions describe training/selection variability on the two fixed panels.

A.4.2 NUMERICAL CHECKS ON LEARNED VIT-L/14 TRANSFORMS

The numerical check uses five selected G states on 12 fixed images, giving 60 image–state comparisons in FP32. For the native path, the learned transform is folded into the last block's Q/K weights and biases before calling the model's ordinary image encoder. The reference is the original image encoder on the same inputs. For the dense path, an independent projection-folding implementation is compared with explicit feature-space application of G, with the same reader and decoder. Table 8 reports maxima over these comparisons.

Table 7: ViT-L/14 absolute mIoU for all five seeds. N is shared within each panel.

Panel	Seed	N	G16	T32	P4
VOC	20260821	42.417	57.847	49.845	64.118
	20260831	42.417	57.858	49.887	63.553
	20260841	42.417	58.157	49.840	64.287
	20260851	42.417	58.110	49.847	65.103
	20260861	42.417	57.880	49.768	65.291
COCO	20260821	19.455	31.528	26.489	31.424
	20260831	19.455	31.031	26.697	31.770
	20260841	19.455	31.790	26.415	30.045
	20260851	19.455	31.365	26.489	30.293
	20260861	19.455	31.442	26.608	30.663

Table 8: Numerical preservation for learned ViT-L/14 transforms. Native errors compare the encoder outputs before and after folding; dense errors compare explicit and independently folded readers. The single mask difference lies at a decision boundary within the measured class-score perturbation.

Quantity	Maximum error or total count
Native image encoding, absolute	3.8147×10^{-6}
Native image encoding, relative Frobenius	3.8398×10^{-7}
Native QK interaction, relative Frobenius	3.8677×10^{-7}
Dense class field, absolute	1.5497×10^{-6}
Dense normalized field, absolute	1.4007×10^{-6}
Different mask pixels across all 60 comparisons	1

Image-encoding errors use the returned encoder features, without an additional normalization step in the audit. Boundary checking compares each changed pixel's top-two or background-threshold margin with its measured field error; no difference falls outside this bound. These checks concern the learned transforms and are distinct from the hand-chosen diagonal checkpoint witness in Section 3.

A.5 Projection and Output-Head Comparisons

A.5.1 Reader-Only Projection Adaptation on ViT-B/16

P4 is the reader-only deployment of the existing full-projection Q/K LoRA4 baseline. It has 13,824 trainable parameters, versus 12,480 for G16. Its dense branch uses the selected Q/K weight and bias updates; its native image–text branch uses the original projections. The dense training loss is unchanged by this path separation. Original, reader-only and independently folded dense implementations agree within the specified field tolerances, with all mask differences explained by decision boundaries. G instead preserves the native computation through its modified Q/K projections.

The full-panel comparison reuses five pairs of development-selected endpoints within a six-epoch candidate budget. Each original arm used four learning rates, observations at epochs 1, 2, 3 and 6, and strengths $\{0.25, 0.5, 0.75, 1\}$ plus identity: 65 candidates per method and seed. G rates were $\{0.002, 0.005, 0.02, 0.08\}$; P4 rates were $\{0.0005, 0.002, 0.005, 0.02\}$. No training or reselection was performed for this evaluation. It completes the historical 1,321-image VOC panel after the earlier 512-image results were known, so its intervals are conditional on this post-hoc analysis rather than a new blinded confirmation.

N, G16 and P4 obtain 58.7821, 62.8228 and 64.5437 mean mIoU. Their paired-seed differences are G$-$P4 $= -1.7208$ [$-2.8599, -0.5817$], G$-$N $= 4.0408$ [3.8129, 4.2686] and P4$-$N $= 5.7616$ [4.6771, 6.8461] pp. These use Student-t intervals with four degrees of freedom and 98.3333% coverage in the three-comparison family. The comparison evaluates complete adaptation recipes rather than isolating reciprocity as a single causal factor.

The full-projection LoRA factors use $W_Q^L = W_Q + B_Q A_Q$ and $W_K^L = W_K + B_K A_K$ in PyTorch's weight convention, with independent Q/K factors and trainable bias increments. The multiplier is one and dropout is zero. At rank four on ViT-B/16, each A is 4×768 and each B is 768×4; A is uniform on $[-1/\sqrt{768}, 1/\sqrt{768}]$, while B and bias increments start at zero. The two studies' G baseline provenance is given in Section D.1.

A.5.2 Standalone Text-Compatible Output Head

The standalone Hout12 applies $Y + s(YA^\top)B^\top$ to 512-dimensional aggregated features immediately before their final L2 normalization and text matching. There is no bias, nonlinearity, dropout or trainable class classifier. A is uniform on $[-1/\sqrt{512}, 1/\sqrt{512}]$ and $B = 0$ at initialization; training uses $s = 1$. Hout12 has 12,288 parameters.

Hout12 and G16 are trained from identity over five seeds for six epochs. Head learning rates are $\{0.0005, 0.002, 0.005, 0.02\}$ and G rates are $\{0.002, 0.005, 0.02, 0.08\}$. Each learning rate is evaluated at epochs 1, 2, 3 and 6 and at four strengths. Adding identity once gives 65 candidates per method and seed. Selection uses dev64 before evaluation on the historical eval1321 panel.

G16 minus Hout12 is -3.894 pp, with a 98.3333% paired-seed interval of $[-5.766, -2.022]$ in the three-comparison family (Table 9).

Table 9: Standalone adaptation. Five paired seeds on eval1321; 98.3333% t intervals for the three-comparison family. N = 58.782, G16 = 62.770, and Hout12 = 66.664 mean mIoU.

Contrast	Mean difference (pp)	Interval
G16 − Hout12	-3.894	[-5.766, -2.022]
G16 − N	+3.988	[3.770, 4.205]
Hout12 − N	+7.882	[6.164, 9.600]

A.6 Adaptation with Fixed Output Heads

The five parents are all selected Hout12 states from the joint-fitting study in Section C.11, whose head-only arms train from identity. Their four learning rates are $\{0.0005, 0.002, 0.005, 0.02\}$ over six epochs; selection observes epochs $\{1, 2, 3, 6\}$ and strengths $\{0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1\}$, plus the unchanged reader (129 candidates per seed). Dev64 pooled mIoU selects the states in Table 10. Their weights and strengths are locked before second-stage training; state hashes and parent–child mappings are in `reproduction.zip`. These parents are distinct from the standalone comparison's selected heads.

Table 10: Selected Hout12 parents reused as H0 in the fixed-head study. Each row is selected on dev64 from the joint-fitting study's head-only arm.

Seed	Learning rate	Selected epoch	Fixed strength s_0
20260821	0.002	6	0.875
20260831	0.002	6	0.875
20260841	0.002	6	0.875
20260851	0.005	6	0.625
20260861	0.005	3	0.625

H0 retains each selected parent. HG freezes its weights and s_0 and trains G16 upstream, preserving gradients through the head. HC continues the parent's factors with s_0 fixed and calibrates the function increment as $H_0 + \alpha(H_\theta - H_0)$. HA freezes the parent and adds a rank-12 residual after its unnormalized output; normalization occurs only after both heads. Each arm uses five paired seeds, four additional epochs and four rates (the same G/head grids above). For each seed, dev64 selects among four strengths $\alpha \in \{0.25, 0.5, 0.75, 1\}$ at each rate/epoch and the unchanged parent, yielding 65 candidates before eval1321 evaluation.

HG−H0 averages 1.971 mIoU points across the five paired seeds, with a 95% t interval of [0.979, 2.962] and positive increments in all five seeds. The secondary comparisons with continued-head training and an added head give 1.771 and 1.167 points, with 97.5% intervals of [0.429, 3.114] and [0.358, 1.975], respectively (Table 11). Neither secondary interval establishes an increment above the prespecified 0.5-point threshold.

Table 11: Metric adaptation with fixed output heads. Five paired seeds on eval1321; primary and secondary interval coverages are reported separately. H0 and HG average 66.476 and 68.447 mIoU.

Contrast	Mean difference (pp)	Coverage	Interval
HG − H0	+1.971	95%	[0.979, 2.962]
HG − HC	+1.771	97.5%	[0.429, 3.114]
HG − HA	+1.167	97.5%	[0.358, 1.975]

A.7 Frozen Reuse Across Data and Readers

A.7.1 The Fixed COCO-Object Panel on ViT-B/16

For this ViT-B/16 confirmation, N scores 30.7313, the mean G score is 33.2899, and mean head-scalar score is 31.6215. The difference is 1.6684 pp [0.9793, 2.2261], using 50,000 paired-image bootstrap draws, PCG64(2026092205), and 99.1667% confidence per contrast in the six-comparison Bonferroni family shared with the shape and bias studies (nominal 95% family-wise coverage). VOC-selected endpoints, temperatures and thresholds are frozen. The later ViT-L/14 study reuses these panel IDs and label conventions, with its own model-specific text embeddings and VOC-selected states. Both studies use 81 class groups and 113 fixed text queries, encoded from class aliases with the same prompt templates using the corresponding backbone's text encoder.

A.7.2 Complete Target-Reader Pipelines

PEARL uses commit `c26062b5062b73f712cbe10846fb3a51e313a6fa` and ResCLIP commit `1f39a42ca1585c4401882c2e3edff10bb7cd5d3b`. We use the complete official pipelines. For PEARL, the scalar multiplies the combined PA and 0.08 centered-KK term before softmax; PA, KK, and TLP are recomputed for each endpoint. For ResCLIP, the scalar multiplies final-layer KK logits before SFR mixing and addition of the spatial prior; intermediate QK, feedback, and PAMR remain part of the computation. These scalars are not the same intervention in the two readers. The chosen baseline scalar is 1 for both; the G16/G8 scalar is 2 for PEARL and 1 for ResCLIP. Thus this is frozen-G reuse with target-reader development calibration, not completely uncalibrated reuse.

For the ResCLIP cross-reader comparison, both the no-G baseline and frozen G16 use the same dev64 images and scalar grid $\{1/8, 1/4, 1/2, 1, 2, 4, 8\}$. The baseline selects its scalar by pooled development mIoU; G16 selects one shared scalar by the mean pooled mIoU of its three fixed source seeds. Ties favor the scalar closest to one on the $\log_2$ scale, then the smaller scalar, and both selected scalars are one. Each group therefore has seven scalar choices, with one baseline evaluation or three seed evaluations per choice, and selection is locked before eval1321.

External reuse instead deploys the three previously frozen rank-8 matrices at $s = 1$ in complete ResCLIP on COCO-Object and ADE20K. No target G optimization, matrix averaging, or seed-prediction fusion is used. Each image is evaluated with seven no-G scalars and three fixed G states. The seven-point target-label temperature oracle selects $s = 1$ for every reporting group. It is a diagnostic oracle, not a target-label-free selector. Its conservative interval uses the Bonferroni-adjusted comparison family covering scalar selection. The non-VOC-60 IoUs include false positives from every valid pixel, including background and other classes, before averaging those fixed 60 labels. They are not computed by cropping evaluation down to non-VOC ground truth. Non-VOC does not mean unseen during CLIP pretraining.

Table 12: Two distinct reuse studies. Upper panel: frozen NACLIP-trained rank-16 G, with separate development selection of target-reader scalar calibration; 99.1667% image intervals (six comparisons). Lower panel: three earlier frozen rank-8 G states in complete ResCLIP at $s = 1$, without target fitting or selection; 99.7619% image intervals (21 comparisons). These are different endpoints and protocols.

Study	Target	Increment	Interval
Cross-reader VOC	ResCLIP: C16−T	+2.455	[2.023, 2.868]
	PEARL: C16−T	+0.848	[−0.009, 1.639]
External ResCLIP	COCO, all 81 classes: G−N	+0.485	[0.346, 0.618]
	COCO, non-VOC 60: G−N	+0.221	[0.060, 0.372]
	ADE, all 150: G−N	+0.465	[0.314, 0.618]

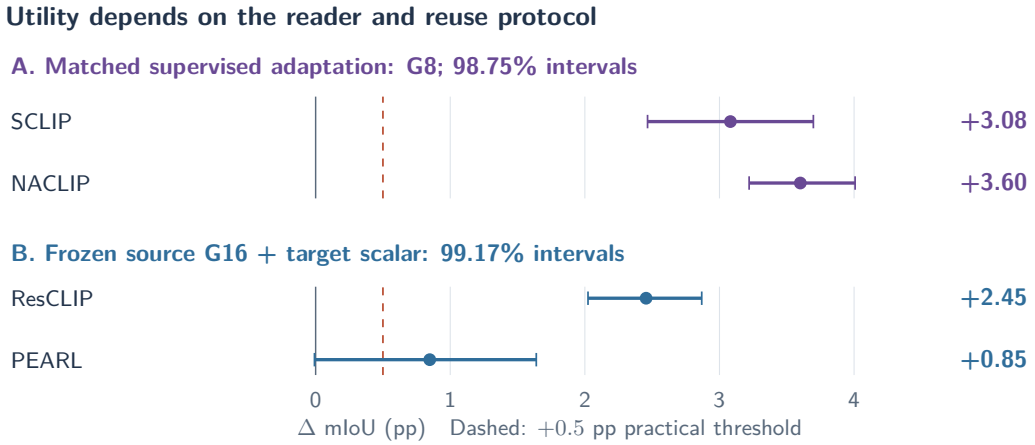

Figure 4: G's utility under two distinct VOC protocols. All contrasts use each reader's development-selected scalar-only baseline. (A) Host-specific supervised G8 fitting, TG−T; 98.75% intervals from a four-contrast family. (B) Frozen NACLIP-trained G16 reused in complete target pipelines with target-reader scalar selection, C16−T; 99.1667% intervals from a six-contrast family. Both use paired-image resampling on 1,321 images, conditional on three selected endpoints in each study.

A.7.3 Frozen Reuse Has Reader- and Target-Dependent Magnitude

Complete ResCLIP improves over its selected scalar baseline. Complete PEARL has three positive seed differences, but its interval includes zero, leaving the increment unresolved.

External reuse gives small positive gains on COCO-Object and ADE. For the non-VOC group, the upper interval bound is 0.372 points, below the prespecified 0.5-point threshold. This group scores 60 classes on the same COCO images, which can also contain VOC classes.

B Proofs, Reader Conditions and Task Transport

B.1 Additional Symmetry and Consistency Results

Consider one attention head with $X \in \mathbb{R}^{n\times m}$, $W_Q, W_K \in \mathbb{R}^{m\times d}$ and $b_Q, b_K \in \mathbb{R}^d$. Including projection biases,

$$Q = XW_Q + \mathbf{1}b_Q^\top, \qquad K = XW_K + \mathbf{1}b_K^\top. \tag{13}$$

The native attention probabilities are $P = \mathrm{softmax}(QK^\top/\sqrt{d})$.

Definition 1 (Functional equivalence)**.** *Two parameter settings θ and θ' are functionally equivalent, written $\theta \sim_F \theta'$, if their native model outputs agree for every input: $F_\theta(x) = F_{\theta'}(x)$.*

Definition 2 (Function-identifiable readout)**.** *A dense readout D is function-identifiable if $\theta \sim_F \theta'$ implies $D_\theta(x) = D_{\theta'}(x)$ for every input x. Equivalently, D must be constant on each model-function equivalence class.*

Definition 3 (Consistency under a function-preserving group)**.** *Let $\mathcal{G}$ act on a parameter domain Θ_0 such that $F_{g\cdot\theta} = F_\theta$ for every $g \in \mathcal{G}$ and $\theta \in \Theta_0$. A readout is $\mathcal{G}$-consistent if $D_{g\cdot\theta}(x) = D_\theta(x)$ for every g, θ, and x.*

The amount of gauge fixing required is consumer-relative. Let $C : \Theta_0 \to \mathcal{Z}$ be a repair and define its residual fiber $\mathcal{F}_C(\theta) = \{C(g \cdot \theta) : g \in \mathcal{G}\}$. Suppose these fibers are exactly the orbits of a residual group $\mathcal{K}$ acting on the left on the invariant admitted image $\mathcal{Z} = C(\Theta_0)$. Here $R(z)$ denotes the entire readout function on the declared input domain, not its output on one sample. Define its uniform stabilizer

$$\mathcal{H}_R = \{k \in \mathcal{K} : R(k \cdot z) = R(z) \text{ for every admitted } z\}. \tag{14}$$

Proposition 3 (Standard residual-invariance criterion)**.** *Under the residual-fiber assumption above, $R \circ C$ is $\mathcal{G}$-consistent if and only if $\mathcal{H}_R = \mathcal{K}$. More generally, for a refinement with an invariant admitted image $\mathcal{Z}'$ and exact residual $\mathcal{L}$-orbits, consistency holds if and only if every $\ell \in \mathcal{L}$ fixes R on all of $\mathcal{Z}'$. Uniform equivalence of transformations, defined by $R(k_1 z) = R(k_2 z)$ for every $z \in \mathcal{Z}$, is represented by $\mathcal{H}_R \backslash \mathcal{K}$. This does not classify all possible single-state output collisions or require a unique full coordinate frame.*

This criterion formulates the partial-canonicalization principle (Kaba et al., 2023, Section 3.3 and Theorem 3.1) with the residual-fiber and domain assumptions needed by our constructions. Section B.2 proves the residual criterion; Section B.3 gives its affine-affinity specialization.

Global function identifiability implies consistency under every such group. Therefore one orbit witness with different dense outputs disproves global identifiability. The converse holds only when the relevant functional-equivalence classes are exactly the group orbits; we do not assume this for a complete Transformer. This distinction lets us use a tractable attention symmetry as a necessary test without claiming to characterize every model symmetry.

Proposition 4 (Function-preserving attention symmetries)**.** *Following established Transformer Q/K and V/O gauge symmetries (Wang & Wang, 2026b), for one head let $Y = PVO$. For $G \in \mathrm{GL}(d)$, $J \in \mathrm{GL}(d_v)$, and $c \in \mathbb{R}^d$, define*

$$Q' = QG, \quad K' = KG^{-\top} + \mathbf{1}c^\top, \quad V' = VJ, \quad O' = J^{-1}O. \tag{15}$$

Then $P' = P$ and $Y' = Y$. Thus reciprocal Q/K changes of basis, a common translation of every key, and reciprocal V/O changes of basis preserve the native head and the complete model function when all other modules are fixed. The common-key translation is the elementary softmax extension used in our audit.

Indeed, $Q'K'^\top = QK^\top + (QGc)\mathbf{1}^\top$ adds a constant within each softmax row, while $V'O' = VO$. Hence $P' = P$ and $Y' = Y$.

For the pure reciprocal Q/K subgroup ($c = 0$, with V/O fixed),

$$Q'Q'^\top = QGG^\top Q^\top, \tag{16}$$

$$K'K'^\top = K(GG^\top)^{-1}K^\top. \tag{17}$$

These raw self-Gram maps are invariant for arbitrary full-rank activations only when G is orthogonal. Key translation additionally changes $KK^\top$, and a V/O gauge can change $VV^\top$. This does not imply that every function that mentions a Gram matrix is non-identifiable: it may retain only an invariant of that matrix. A particular dense readout is refuted by an explicit orbit witness; our certificate-conditioned experiments provide such witnesses for current self-correlation methods.

The proposition separates an interaction from a metric. Native attention identifies a query–key pairing only modulo row-wise logit translations; it does not identify an intrinsic Euclidean metric within query or key coordinates, nor a unique logit matrix. A semantic claim based on a latent metric therefore needs an additional convention or a construction consistent with the audited symmetry.

Let $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ center the key-token axis. The pair quotient also has a necessary-and-sufficient activation-level form.

Proposition 5 (Maximal invariant of an affine Q/K pair)**.** *Assume Q and HK have full column rank. Under $Q \mapsto QG$ and $K \mapsto KG^{-\top} + \mathbf{1}c^\top$, the centered interaction*

$$M_c = QK^\top H \tag{18}$$

is a maximal invariant: two pairs have equal M_c if and only if they are related by one such affine Q/K action. Consequently, every readout of the pair that is invariant under this full-rank affine action factors through M_c.

Thus an intrinsic affine-Q/K readout may use the interaction pattern exposed by native attention, but cannot assign a unique Euclidean metric to either factor without an additional convention. This orbit characterization is a structural corollary of full-rank matrix-factorization uniqueness. The affine activation-pair argument in Section B.2 gives the proof.

Proposition 6 (Maximal invariant of one attention factor)**.** *Let $Q \in \mathbb{R}^{n\times d}$ have full column rank and let $\mathrm{GL}(d)$ act by $Q \mapsto QG$. The column-space projector*

$$\Pi_Q = Q(Q^\top Q)^{-1}Q^\top \tag{19}$$

is a maximal invariant: $\Pi_Q = \Pi_{Q'}$ if and only if $Q' = QG$ for a unique $G \in \mathrm{GL}(d)$. Consequently, every single-factor readout that is invariant to the full right action factors through Π_Q. The same statement holds for K.

Equality of the projectors is equivalent to equality of column spaces. With full column rank, this gives the unique invertible change of basis $Q' = QG$.

The invariant constrains recovery of coordinate-dependent targets. Consider a checkpoint-native target $T(Q)$ derived from the Euclidean self-metric, such as a Top-k graph of $QQ^\top$. If an orbit pair Q and QG has $T(Q) \neq T(QG)$, then any invariant prediction R satisfies

$$\max\{d(R(Q), T(Q)), d(R(QG), T(QG))\} \geq \tfrac{1}{2}d(T(Q), T(QG)) \tag{20}$$

for every metric d on target outputs. Thus a readout cannot simultaneously be single-factor, full-GL invariant, and universally faithful to checkpoint-native Euclidean neighborhood semantics. This does not preclude a useful invariant readout of ground-truth semantics, nor a construction using the interaction $QK^\top$ (or $QK^\top H$ for the affine action) or an explicit gauge convention. Since $R(Q) = R(QG)$, Eq. (20) follows directly from the triangle inequality.

Predicted taxonomy. Under the reciprocal Q/K subgroup used in our main audit, standard $QK^\top$ attention, a fixed spatial kernel, and a value-only affinity are invariant because V is held fixed. Raw $QQ^\top$, $KK^\top$, and norm-scaled spatial rules are generally sensitive. Value-only affinity is not globally identifiable under the independent V/O gauge. These subgroup-scoped predictions are fixed before inspecting downstream scores and serve as implementation controls.

B.2 Affine Canonicalization: Proof and Reader Scope

Consumer-relative residual resolution. For a repair C, fix θ and consider the set $\mathcal{F}_C(\theta) = \{C(g \cdot \theta) : g \in \mathcal{G}\}$. By assumption this set equals $\mathcal{K} \cdot C(\theta)$. Therefore $R \circ C$ takes the same value on the complete $\mathcal{G}$-orbit of θ if and only if $R(k \cdot C(\theta)) = R(C(\theta))$ for every reachable $k \in \mathcal{K}$. Requiring this for every admitted θ is exactly $\mathcal{H}_R = \mathcal{K}$, where $R(z)$ is the entire readout function. For a refinement, apply the same proof on its own invariant admitted image $\mathcal{Z}'$ and its exact residual $\mathcal{L}$-orbits; the stabilizer must be defined on that image.

To check the coset convention, define $k_1 \equiv_R k_2$ by $R(k_1 z) = R(k_2 z)$ for every $z \in \mathcal{Z}$. Domain invariance and the substitution $w = k_1 z$ give $k_1 \equiv_R k_2 \iff k_2 k_1^{-1} \in \mathcal{H}_R$. The set $\mathcal{H}_R$ is a subgroup: identity fixes R, products preserve it by composition, and inverse invariance follows by replacing z with $h^{-1}z$. Hence the equivalence classes are $\mathcal{H}_R k$, giving $\mathcal{H}_R\backslash\mathcal{K}$ for the declared left action. It is a quotient group only if $\mathcal{H}_R$ is normal. These are equivalence classes of transformations acting on the entire admitted domain, not all fibers of R on individual states. For example, with $R(z) = z_1$ on $\mathbb{R}^2$ and $\mathcal{K} = \mathrm{O}(2)$, let $h = \mathrm{diag}(1, -1)$ and let k swap coordinates. Then k and kh lie in the same coset $k\mathcal{H}_R$ but $R(ke_2) = 1$ and $R(khe_2) = -1$. Without exact reachable residual fibers, invariance to a declared residual group remains sufficient; necessity applies only to realized residual pairs.

Orbit characterization. Following the established Q/K gauge and quotient geometry (Wang & Wang, 2026b;a), let A, C and A', C' have full column rank d. Under the action $(A, C) \mapsto (AG, CG^{-\top})$, the cross-operator is a complete orbit invariant:

$$AC^\top = A'C'^\top \quad \Longleftrightarrow \quad (A', C') = (AG, CG^{-\top}) \tag{21}$$

for a unique $G \in \mathrm{GL}(d)$. Indeed, the common rank-d product gives $\mathrm{col}(A) = \mathrm{col}(A')$, hence $A' = AG$ for a unique invertible G; applying a left inverse of A then gives $C' = CG^{-\top}$. Thus full-rank Q/K pairs modulo $\mathrm{GL}(d)$ are characterized by their identifiable cross-operator. This is a per-head statement and does not characterize a complete Transformer's functional-equivalence classes.

Affine activation-pair maximal invariant. Let $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. Under $Q' = QG$ and $K' = KG^{-\top} + \mathbf{1}c^\top$,

$$Q'K'^\top H = QG(G^{-1}K^\top + c\mathbf{1}^\top)H = QK^\top H, \tag{22}$$

because $\mathbf{1}^\top H = 0$. Conversely, suppose $QK^\top H = Q'K'^\top H$ and both Q, HK, Q', HK' have full column rank. Applying the preceding full-rank factorization result to $Q(HK)^\top = Q'(HK')^\top$ gives a unique G with $Q' = QG$ and $HK' = HKG^{-\top}$. Therefore $H(K' - KG^{-\top}) = 0$, so every column of the difference is constant over tokens: $K' = KG^{-\top} + \mathbf{1}c^\top$ for a unique c. Equality of centered interactions therefore characterizes exactly one affine Q/K orbit. Any invariant of the pair factors through this centered operator by the usual maximal-invariant construction. The result concerns this activation-level action; it does not claim that the centered operator completely identifies a whole Transformer's function.

Balanced quotient. GaugeKV proposes Q/K Gram balancing (Wang & Wang, 2026c); the QR partial slice of Wang & Wang (2026b) uses a different representative. We derive the positive-definite balancing form independently, fix key translation, and give its dense-readout consequence. Let $A = W_Q$ and $C = W_K$ have full column rank d. Then $S_A = A^\top A$ and $S_C = C^\top C$ are positive definite. The equation $PS_AP = S_C$ has a unique positive-definite solution: setting $Y = S_A^{1/2} P S_A^{1/2}$ gives $Y^2 = S_A^{1/2} S_C S_A^{1/2}$, whose positive-definite square root is unique. Therefore

$$P = S_A^{-1/2}(S_A^{1/2} S_C S_A^{1/2})^{1/2} S_A^{-1/2} \tag{23}$$

is uniquely defined, as is its positive-definite square root $B = P^{1/2}$.

Define the canonical factors $\bar{A} = AB$ and $\bar{C} = CB^{-\top}$. Because B is symmetric positive definite,

$$\bar{A}^\top \bar{A} = BS_AB, \tag{24}$$

$$\bar{C}^\top \bar{C} = B^{-1}S_CB^{-1} = B^{-1}(B^2S_AB^2)B^{-1} = BS_AB. \tag{25}$$

Thus both canonical factors have the same positive-definite Gram Γ, while $\bar{A}\bar{C}^\top = AC^\top$.

Now take any $G \in \mathrm{GL}(d)$, form $A' = AG$, $C' = CG^{-\top}$, and canonicalize this representative independently to $(\bar{A}', \bar{C}')$. Both canonical pairs factor the same rank-d operator $AC^\top$. Their column spaces therefore agree, so there is a unique invertible R such that $\bar{A}' = \bar{A}R$. Equality of the cross-products and a left inverse of $\bar{A}$ then give $\bar{C}' = \bar{C}R^{-\top}$. Since each pair has equal query and key Grams,

$$R^\top \Gamma R = R^{-1}\Gamma R^{-\top}. \tag{26}$$

Multiplying on the left by R and on the right by $R^\top$ yields $X\Gamma X = \Gamma$ for the positive-definite matrix $X = RR^\top$. Let $Z = \Gamma^{1/2} X \Gamma^{1/2}$. Then $Z^2 = \Gamma^2$; uniqueness of the positive-definite square root gives $Z = \Gamma$, hence $X = I$. Therefore R is orthogonal and $R^{-\top} = R$. The two canonical pairs differ by one shared orthogonal transform, so both self-Grams are identical.

Reciprocally transformed biases obey the same linear result. For token matrix T, write $Q = TA + \mathbf{1}b_Q^\top$ and $K = TC + \mathbf{1}b_K^\top$. Under a gauge, $b_Q' = G^\top b_Q$ and $b_K' = G^{-1}b_K$, giving $Q' = QG$ and $K' = KG^{-\top}$. Canonicalization applies the same right actions B and $B^{-\top}$ to weights and affine rows. The factor relation above therefore gives $\bar{Q}' = \bar{Q}R$ and $\bar{K}' = \bar{K}R$. Consequently any dense readout whose dependence on the canonical Q/K coordinates is invariant to a shared orthogonal transform – including $QQ^\top$, $KK^\top$, their mixtures, and downstream operations on those token Grams

– is constant over the full per-head Q/K gauge orbit. The proof does not cover rank-deficient factors or readouts that separately use non-orthogonally invariant Q/K coordinates.

The additional key-translation action is $b'_K = G^{-1}b_K + c$. Setting b_K to zero before linear balancing maps every such translated representative to the same linear orbit while preserving native softmax attention. After canonicalization, head h retains an independent shared $R_h \in \mathrm{O}(d_h)$; the residual group is therefore $\prod_h \mathrm{O}(d_h)$, not one global rotation. A head-index-sensitive readout would additionally need to respect any allowed head permutation. The affine construction does not address the independent V/O gauge or any whole-network symmetry.

The common balanced Gram. Wang & Wang (2026c, Proposition 2.9 and Appendix E.1) identify the common balanced Gram with $S_A \# S_C$. This equality fails for general noncommuting Grams. For example, take

$$S_A = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}, \quad S_C = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}, \quad F = S_A \# S_C = \frac{1}{\sqrt{37}} \begin{bmatrix} 22 & 8 \\ 8 & 13 \end{bmatrix}. \tag{27}$$

If some invertible D satisfied $D^\top S_A D = D^{-1} S_C D^{-\top} = F$, then $F^2 = D^\top S_A S_C D^{-\top}$, requiring $\mathrm{tr}(F^2) = \mathrm{tr}(S_A S_C)$. Instead these traces are $781/37$ and 25. Thus the matrix geometric mean need not be the common balanced Gram. The SPD-equation construction above gives the balanced factors directly.

Method-level orthogonal-invariance corollary. Let two canonicalized executions satisfy $Q' = QR_h$ and $K' = KR_h$ in every head, with R_h orthogonal and all other inputs fixed. Then $Q'Q'^\top = QQ^\top$, $K'K'^\top = KK^\top$, and $Q'K'^\top = QK^\top$ head by head. All deterministic operations downstream of these matrices therefore agree. This directly covers the matched SCLIP and NACLIP formulas. It also covers ResCLIP because its intermediate RCS route uses unchanged earlier-layer QK matrices while the changed final route consumes the invariant self-correlation.

For PEARL PA+KK we prove the implemented finite iteration directly. Suppress the head index and write its norm-derived token weights as $w_i = \|q_i\|/(\sum_j \|q_j\| + \epsilon)$, with the CLS weight zeroed before normalization. Under $Q' = QR, K' = KR$, orthogonality makes every w_i invariant; the weighted means and centered arrays transform on the right by R. Thus $M = K_c^\top Q_c$ transforms as $M' = R^\top M R$. The pinned `polar` path uses $\epsilon = 10^{-6}$ and five iterations:

$$A = M^\top M, \qquad a = \max(\|A\|_F, \epsilon), \qquad Y_0 = A/a + \epsilon I, \qquad Z_0 = I, \tag{28}$$

$$T_t = \tfrac{1}{2}(3I - Z_t Y_t), \quad Y_{t+1} = Y_t T_t, \qquad Z_{t+1} = T_t Z_t, \qquad U_5(M) = M Z_5/\sqrt{a}. \tag{29}$$

The Frobenius norm and clamp are invariant under orthogonal conjugation. By induction A, Y_t, Z_t, T_t all conjugate by R; therefore $U_5(M') = R^\top U_5(M) R$, even when M is singular or zero. Since $K' U_5(M') = K U_5(M) R$, both the aligned QK logits and the centered KK term are invariant. Fixed V, output projection, and deterministic downstream operations then preserve the pre-TLP output. The code's gradient detach changes no forward value. This is finite-step equivariance in exact arithmetic, not convergence to an orthogonal optimizer: for $M = 0$ the returned map is zero. A 50-token B/32 crop has $\mathrm{rank}(M) \le 50 < 64$, so an invertible-covariance proof would not cover it. This activation rank limitation does not violate the independent full-rank assumption on projection weights in the CAQC theorem. CPU tests cover singular/zero inputs and the pinned full PA+KK path; floating-point casts yield tolerance-based, not bitwise, invariance. The alternate SVD path needs a unique polar factor or an equivariant choice at singularities and is not admitted by this finite-iteration proof.

Table 13: Dense-reader inputs preserved by the shared orthogonal action. The result covers the listed implementations; for PEARL it covers PA+KK before TLP.

Readout	Role of similarity	Gauge-relevant inputs	Implementation covered
SCLIP	patch self-affinity	$QQ^\top, KK^\top$	matched formula
NACLIP	K similarity as patch content	$KK^\top$ + spatial prior	matched + official wrapper
ResCLIP	self-correlation corrected by earlier QK	final self-correlation + earlier QK	pinned official wrapper
PEARL	orthogonal Q/K basis repair	aligned QK + $KK^\top$	PA+KK before TLP

B.3 Exact Residual Condition for an Affine Self-Affinity

The following elementary specialization makes the consumer condition explicit. It uses the row-token convention and concerns one attention branch. Let $Z \in \mathbb{R}^{n\times d}$, $M \succ 0$, $\beta \in \mathbb{R}^d$, and let $\Omega \in \mathbb{R}^{n\times n}$ have finite entries. Define

$$L_{M,\beta}(Z) = ZMZ^\top + \mathbf{1}\beta^\top Z^\top, \qquad A_{M,\beta}(Z) = \operatorname{softmax}\left(L_{M,\beta}(Z)/\sqrt{d} + \Omega\right). \tag{30}$$

For $Y = ZB + \mathbf{1}b^\top$ with invertible B, its affine self-Gram gives the same attention with $M = BB^\top$ and $\beta = Bb$: the omitted terms $ZBb\mathbf{1}^\top + \|b\|^2\mathbf{1}\mathbf{1}^\top$ are constant within each row. Here β is the effective bias vector.

Proposition 7 (Residual compatibility of an affine self-affinity)**.** *For $O \in \mathrm{O}(d)$, put $\Delta_M = OMO^\top - M$, $\delta_\beta = O\beta - \beta$, and $H_n = I - \mathbf{1}\mathbf{1}^\top/n$. Then*

$$A_{M,\beta}(ZO) = A_{M,\beta}(Z) \quad \Longleftrightarrow \quad \left[Z\Delta_M + \mathbf{1}\delta_\beta^\top\right](H_nZ)^\top = 0. \tag{31}$$

If $\operatorname{rank}(H_nZ) = d$, this is equivalent to $OMO^\top = M$ and $O\beta = \beta$. At any feature rank, the transported pair $M' = O^\top MO$, $\beta' = O^\top\beta$ satisfies $A_{M',\beta'}(ZO) = A_{M,\beta}(Z)$.

Proof. Two finite logit arrays have equal row-softmax if and only if their difference is constant within each row: this follows by taking log ratios of their positive probabilities. Thus equality holds exactly when the logit difference right-multiplied by H_n is zero. The common Ω cancels, and expanding the difference gives (31). If H_nZ has full column rank, so does $[Z, \mathbf{1}]$: apply H_n to any linear dependence to eliminate its constant column. Factor the right side as

$$[Z, \mathbf{1}] \begin{bmatrix} \Delta_M \\ \delta_\beta^\top \end{bmatrix} (H_nZ)^\top. \tag{32}$$

A left inverse of $[Z, \mathbf{1}]$ and a right inverse of $(H_nZ)^\top$ give $\Delta_M = 0$ and $\delta_\beta = 0$. The converse and the transported identity follow by substitution and $OO^\top = I$. □

What needs a frame. For $\beta = 0$, the remaining symmetry is the orthogonal stabilizer of M. Writing $M = U\operatorname{diag}(\lambda_1 I_{m_1}, \ldots, \lambda_k I_{m_k})U^\top$ with distinct λ_j, this stabilizer is $U(\prod_j \mathrm{O}(m_j))U^\top$. This is the standard spectral stabilizer structure (cf. the $\mathrm{SO}(d)$ version in Groisser et al., 2017, Section 2 and Appendix A): rotations within an equal-eigenvalue block do not change the metric. With nonzero β, each block must also fix the corresponding component of $U^\top\beta$. Consequently $M = cI, \beta = 0$ needs no residual frame resolution, while an isotropic metric with fixed nonzero bias can already be sensitive. APPF resolves a full frame, which is sufficient but can fix more coordinates than a particular readout requires.

Rank and output boundaries. The necessity above is an activation-rank statement, not the projection-weight rank assumption of CAQC. Full rank of Z alone is insufficient for the joint metric-and-bias conclusion. For example, take rows $z_i = (x_i, 1)$ with nonconstant x_i, and

$$M = \begin{bmatrix} 2 & c \\ c & 1 \end{bmatrix}, \quad \beta = (-c, 0)^\top, \quad O = \operatorname{diag}(-1, 1), \quad 0 < |c| < 1. \tag{33}$$

Here both the metric and bias change under O, but the logit difference is $-2cx\mathbf{1}^\top$, so attention is unchanged. For multiple attention branches, preservation of each branch is sufficient, not generally necessary for their sum. Even if attention changes, $(A' - A)V$ can vanish, for example when all value rows are equal. Subsequent hard classification can identify still more outputs. These examples separate affinity invariance from equality of the final masks. The identities apply to fixed finite Ω; hard-masked attention supports and other coordinate-consuming paths need separate analysis.

B.4 Certificates and Task-Endpoint Transport

Reader conditions and numerical checks. CAQC requires the reader to be invariant under the residual group $\mathcal{K} = \prod_h \mathrm{O}(d_h)$. By Proposition 3, this is equivalent to its uniform stabilizer containing $\mathcal{K}$; a globally unique frame is unnecessary. Per-head $QQ^\top$, $KK^\top$, $QK^\top$, deterministic mixtures of these matrices, and equivariant Procrustes alignment satisfy this condition. Readers using raw Q/K coordinates or cross-head coordinate couplings require a separate invariance proof.

Before applying the rewrite, we check full rank, positive definiteness, conditioning, the balanced-Gram residual, idempotence and finite candidate tensors. Supplied probe tokens support numerical comparisons of centered QK, native attention and model outputs. These finite checks validate the numerical implementation; reader invariance is established analytically. If a check fails or valid probes are unavailable, the source parameters are left unchanged.

Offline implementation. The one-time packed-parameter rewrite takes 157 ms on ViT-B/16, adds no inference operators, and keeps checkpoint and stress errors below 7.5×10^{-6}. All checks occur during this offline rewrite.

B.4.1 Compiling One Task Endpoint Across Equivalent Sources

CAQC intentionally stops at a residual shared orthogonal action per head. This is sufficient for its admitted invariant consumers, but a fixed anisotropic Learned G generally does not commute with that action. Consequently, applying the same G after independently canonicalizing equivalent sources can reintroduce source-dependent dense outputs.

This is the consumer-dependent transport case described by Proposition 3. A self-metric induced by $A = GG^\top$ has orthogonal stabilizer $\mathcal{H}_A = \{O : OAO^\top = A\}$. Isotropic A leaves the full orthogonal group invisible, whereas an anisotropic A has a smaller stabilizer and therefore requires finer transport. CTGC chooses a single endpoint, so it resolves the entire remaining frame rather than only the cosets needed by one downstream statistic.

We resolve the remaining orientation using an *Ambient Pivot–Polar Frame* (APPF). Write a source's canonical projections relative to canonical reference 0 as $\bar{W}_{Q,s} = \bar{W}_{Q,0}R_s$ and $\bar{W}_{K,s} = \bar{W}_{K,0}R_s$, with $R_s \in \mathrm{O}(d)$. The query-row matrix is $X_s = \bar{W}_{Q,s}^\top = R_s^\top X_0$. Pivoted Gram–Schmidt selects a nonsingular ambient-column block C_s using residual column-norm scores. Away from pivot ties, the scores and pivot indices are invariant, so $C_s = R_s^\top C_0$. Its right polar decomposition satisfies $C_s = U_s P_0$ with $U_s = R_s^\top U_0$. Consequently

$$\bar{W}_{Q,s}U_s = \bar{W}_{Q,0}U_0, \qquad \bar{W}_{K,s}U_s = \bar{W}_{K,0}U_0.$$

Let B_0 balance the original reference projections, and let G define the frozen task endpoint. The compact artifact stores $H = (B_0U_0)^{-1}G$. Folding it after the source frame recovers

$$\bar{W}_{Q,s}U_sH = W_{Q,0}B_0U_0H = W_{Q,0}G,$$
$$\bar{W}_{K,s}U_sH^{-\top} = W_{K,0}B_0^{-\top}U_0H^{-\top} = W_{K,0}G^{-\top}.$$

The query-bias row follows the same right action. Restoring the stored target key bias $b_{K,\star} = G^{-1}b_{K,0}$ completes the affine endpoint. The artifact also stores pivot indices, thresholds and provenance hashes.

The resulting *Certified Task-Aware Gauge Compiler* (CTGC) runs CAQC and APPF on a source, folds H into Q/K, restores the target affine bias convention, and checks frame admissibility, reference pivot agreement, finite candidate tensors, and native-output parity on supplied probe tokens. If a check fails or probes are missing or invalid, the source parameters are left unchanged. Offline validation

additionally measures endpoint proximity and idempotence, and verifies that non-Q/K tensors remain unchanged. The compiler accepts all five prespecified affine sources for the selected rank-8 G endpoint (6,240 optimization parameters); maximum endpoint relative L_∞ error is 6.754×10^{-7}, native-output drift is 9.537×10^{-7}, and idempotence error is 7.368×10^{-7}. Non-Q/K tensors remain byte-identical. The naive fixed-task-artifact control reaches 1.1013 relative parameter error. It includes reference-balancing compensation and the target key bias, rather than an unadjusted G appended to CAQC. This measures endpoint incompatibility.

CTGC adds no inference operator and transports a fixed supervised endpoint on the admitted generic affine stratum. Exactly identical endpoints inherit their task behavior. With finite-precision parameter differences, hard-mask preservation also depends on decision margins; Section C.12 measures the resulting predictions. Ordinary CAQC remains the reference-free consistency repair and the fallback whenever CTGC abstains.

B.5 Cross-Checkpoint APPF Admission and Stability

We examine APPF stability across six pretrained checkpoints from OpenAI, OpenCLIP/LAION and MetaCLIP, covering B/16, B/32 and L/14. For every numerical condition and head, we measure Q/K rank, minimum relative pivot gap and selected block condition number. Five deterministic affine gauges test pivot agreement, native-function parity and endpoint recovery. The numerical conditions are FP32, FP16 roundtrip, per-output-row symmetric INT8 fake quantization/dequantization, and relative Gaussian perturbations at 10^{-6} and 10^{-4}.

Table 14: FP32 APPF admission across real checkpoints. The admission threshold is 10^{-6}; the prespecified clear-margin threshold is 10^{-5}. All selected block condition numbers are below the 10^6 limit.

Checkpoint	Heads	Minimum gap	Maximum condition	Admitted
OpenAI B/16	12	2.1604×10^{-5}	16.300	yes
OpenAI B/32	12	5.2931×10^{-6}	14.258	yes
OpenAI L/14	16	1.1962×10^{-5}	22.516	yes
OpenCLIP LAION-400M B/16	12	4.1029×10^{-5}	21.477	yes
MetaCLIP-400M B/16	12	4.0858×10^{-6}	18.028	yes
MetaCLIP-fullcc B/16	12	9.5050×10^{-6}	16.438	yes

All 76 FP32 heads are full-rank and all exact-gauge pivot sequences agree. All six checkpoints meet the admission threshold, while three meet the stronger clear-margin criterion.

Table 15: Numerical APPF stability relative to the FP32 pivot sequence. Each row contains 76 heads and six checkpoint-conditions. INT8 changes the numerical representative substantially; it is included as a coverage stress test, not as a native-function-preserving gauge.

Condition	Pivot agreement	Admitted checkpoints	Head abstentions
FP32	76/76 (100.000%)	6/6	0/76
FP16 roundtrip	69/76 (90.789%)	6/6	0/76
Per-row symmetric INT8	1/76 (1.316%)	6/6	0/76
Relative Gaussian 10^{-6}	76/76 (100.000%)	6/6	0/76
Relative Gaussian 10^{-4}	72/76 (94.737%)	5/6	1/76

The only admission failure occurs for OpenAI L/14 under relative 10^{-4} noise, where one head's gap falls to 6.727×10^{-7}. Four of the five equivalent affine sources fall below the admissible pivot-gap threshold, leaving 146/150 sources accepted overall. Across all five numerical conditions, one of six checkpoints, one of 30 checkpoint-conditions, and one of 380 head-conditions abstains at least once; all FP32 abstention rates are zero. Among accepted sources, endpoint relative L_∞ error is at most 1.088×10^{-6} and native-output drift at most 1.907×10^{-6}. Across all sources, centered-QK, attention-probability, and native-output drifts are at most 1.311×10^{-6}, 1.490×10^{-7}, and 1.907×10^{-6}, respectively.

The single-anchor frame admits all six checkpoints in FP32. Its coverage narrows under the stronger Gaussian perturbation, where four sources fail the pivot-gap threshold.

C Additional Evidence and Alternative Designs

C.1 Equivalent-Model Evaluation Protocol

These earlier studies pair a native-function certificate with dense-output measurements. Micro pixel disagreement is the primary consistency statistic, not a measure of semantic quality.

Checkpoint construction and certificate. We gauge the final visual block of official OpenAI CLIP ViT-B/16, transform packed query/key weights and biases, serialize the full state dictionary, and reload it strictly into a fresh model. We then certify native attention logits, normalized image embeddings, image–text logits, and top-1 predictions on all 50,000 ImageNet-1K validation images (Deng et al., 2009). A dense comparison is eligible only if native QK/logit error is at most 2×10^{-4} with no unexpected prediction change; otherwise the counterfactual is invalid. The all-input guarantee comes from the algebraic action; a finite validation set checks its implementation. For affine key translation we compare centered QK and attention, not uncentered logits.

Gauge families. The audit includes signed permutations and orthogonal transformations as negative controls, determinant-balanced diagonal gauges at multiple condition numbers and seeds, and non-axis-aligned dense SPD and shear gauges. Mild-gauge directions and thresholds are frozen on calibration data before confirmation. The audit tests equivalent parameterizations, not random perturbations of the model function.

We classify readouts algebraically before testing native QK, fixed spatial, MaskCLIP-style value readout, norm-scaled spatial affinity, KK, SCLIP, ClearCLIP, NACLIP, ResCLIP, and Sparse Entmax on VOC2012 (Everingham et al., 2010) and ADE20K (Zhou et al., 2017). Unless stated otherwise, paired intervals resample whole images, recompute aggregate confusion-matrix mIoU, and use 10,000 percentile-bootstrap draws. Conjunctive hypotheses require all prespecified conditions to hold; exploratory pairwise intervals are pointwise, without multiplicity adjustment. Protocol- specific exceptions are described in Section D.2. Historical mIoU scorers omit zero-union classes separately by condition and draw. Section D.2 reports post-hoc sensitivity analyses with common class support, including cohorts with missing classes.

Canonicalization selection and external evaluation. Pilots select full weight-Gram canonicalization over diagonal balancing and reject three activation-based alternatives. Frozen confirmation covers all VOC images, a 1,360-image ViT-B/32–ADE split excluded from method selection, and all 5,000 COCO-Stuff validation images with a new ViT-L/14 backbone. For the latter, we first validate the unmodified NACLIP pipeline, then change only its final visual Q/K parameters. The prespecified criteria require native-function preservation, consistent dense outputs and at most one mIoU point of identity loss. Section C.4 reports the confirmation cohorts and outcomes.

Certificate-conditioned consistency statistic. For eligible checkpoint pairs, let V_i be the valid pixels of image x_i and $y_{D,\theta}(x_i)_p$ the hard label produced by dense readout D. Our primary, label-free statistic is the size-weighted micro pixel disagreement

$$\Delta_{\mathrm{pix}}(D, \theta, g; \mathcal{X}) = \frac{\sum_i \sum_{p \in V_i} \mathbf{1}[y_{D,\theta}(x_i)_p \neq y_{D,g(\theta)}(x_i)_p]}{\sum_i |V_i|}. \tag{34}$$

This statistic uses no class-label values conditional on V_i; our benchmark validity masks do use the annotation's ignore mask. Thus "label-free" here does not mean that the benchmark mask convention is annotation-free. Gauge-dependent observations are tied to their specified transformations rather than treated as intrinsic task scores. Per-image macro disagreement, class-aware prediction-mask IoU, and probability Jensen–Shannon divergence are secondary diagnostics. Ground-truth mIoU and its image-paired confidence interval remain separate utility consequences, because positive and negative semantic changes can cancel in aggregate accuracy.

C.2 Checkpoint and Readout Controls

C.2.1 Same Native Model, Different Dense Readout

Table 16: Checkpoint-level counterfactual. Native CLIP behavior is identical on all 50,000 ImageNet validation images after serialization and reload, while the KKV Entmax dense readout changes on the first 128 sorted VOC validation IDs.

Native-function endpoint	Difference
ImageNet validation images	50,000
Top-1 prediction flips	0
Maximum normalized-embedding drift	0
Maximum image-text logit drift	0
VOC KKV Entmax mIoU	$34.135 \rightarrow 19.664$
Dense delta	-14.471 pp

The separate factor-$1/2, 2$ checkpoint roundtrip in Table 16 rules out changes in input, training, native attention, image representation, classification logits, and model loading state. Runtime-transformed and reloaded-checkpoint dense features also match exactly, showing that the dense change is carried by a valid serialized parameterization rather than by an instrumentation-only hook.

C.2.2 The Algebra Predicts Which Readouts Change

Table 17: Controlled-reader diagonal-gauge effects on CLIP ViT-B/16: 1,449 VOC and 2,000 ADE images. Each head uses equal numbers of diagonal entries 8 and $1/8$, randomly assigned to coordinates with seed 47. Zeroes are exact within the evaluator.

Readout	Pred.	VOC Δ	ADE Δ
MaskCLIP-style	invariant	0.000	0.000
Native QK	invariant	0.000	0.000
Fixed spatial	invariant	0.000	0.000
Norm-scaled spatial	sensitive	$+2.006$	-2.302
KK	sensitive	-8.440	-3.941
NACLIP	sensitive	-8.780	-3.956
SCLIP	sensitive	-4.888	-2.220
ClearCLIP	sensitive	-2.971	-3.229

Every algebraically invariant family remains unchanged, whereas every predicted sensitive family changes. This control pattern is more informative than the sign of any one score change. The sign can depend on architecture and gauge direction; the failure of function identifiability does not.

Two interventions localize what changes. Pure isotropic scaling failures vanish exactly under analytic temperature compensation. Determinant-one anisotropy still changes 13.4–14.9% of Top-10 Q/K neighbors after scalar calibration, although a task-level effect is not established in both readouts. Separately, injecting only the transformed self-metric exactly reproduces the SCLIP/NACLIP output change, while restoring the native self-metric exactly rescues token outputs, patch logits, and final masks. This identifies the self-metric as the changing path in these fixed implementations. The following section gives the scalar and path-intervention outcomes.

C.3 Calibration and Causal-Path Controls

We separate scalar calibration from anisotropic geometry on 128 frozen VOC images. This earlier study uses 2,000 paired-image draws with seed 20260804, distinct from the newer supervised temperature study. For the isotropic actions $G = 0.5I$ and $G = 2I$, analytically compensating the self-correlation temperature restores SCLIP and NACLIP predictions exactly. For determinant-one diagonal and dense-SPD gauges, development-only scalar temperature, branch-weight, entropy, variance, and label-free JSD matching do not restore the original rank geometry. Top-10 neighbor sets

change by 14.91% in Q and 13.40% in K; mutual-10NN edge disagreement is 25.47% and 24.45%. These changes establish geometry beyond scalar calibration, while a task-level effect in both readouts remains unestablished.

We then intervene on the proposed mediator while freezing every downstream operation. In the identity path, replacing only $QQ^\top/KK^\top$ by the gauged self-metric reproduces the gauged token output, patch logits, and final mask with zero recorded error for both SCLIP and NACLIP. Conversely, restoring the identity self-metric inside the gauged path recovers all three outputs exactly. Injecting only NACLIP's fixed spatial prior produces zero disagreement. These sufficiency and rescue tests identify the changing self-metric as the causal path in the evaluated implementations.

C.4 Canonicalization, Balancing Statistics and Utility

C.4.1 Choice of Balancing Statistic

The balancing object determines the canonical self-metrics. For the same full-column-rank factor pair, SVD and SPD balancing differ only by a common orthogonal action and produce the same self-Grams. The historical constructions instead use different statistics: bias-augmented factors include the query and key offsets in their Grams, whereas Weight-Gram balancing uses the linear weights. A later intervention fixes the reader pipeline and changes which bias terms enter that statistic.

The historical bias-augmented construction balances affine Q/K factors. For each head, weights and biases were augmented into F_Q, F_K. We formed the identifiable checkpoint operator $B = F_Q F_K^\top = U\Sigma V^\top$ and used the canonical factors $F_Q^c = U\Sigma^{1/2}$ and $F_K^c = V\Sigma^{1/2}$ in the otherwise unchanged SCLIP route. Unlike an earlier image-level factorization of $QK^\top$, this construction uses one checkpoint-level factorization for every image.

Factor reconstruction error was 1.49×10^{-7}, canonical/native QK-logit error was 2.10×10^{-5}, and maximum output drift across signed-permutation, positive-diagonal, and dense-SPD gauges was 7.63×10^{-6}. It nevertheless reached only 42.411 mIoU versus 58.010 for SCLIP and 57.965 for diagonal balancing on the frozen 256-image VOC pilot. The paired difference from SCLIP was -15.599 points (95% CI $[-18.588, -13.315]$). The construction did not meet any of the three prespecified utility criteria, so evaluation was limited to the 256-image pilot. No temperature or mixture was tuned on these images.

The balancing statistic affects task utility. In the legacy VOC-256 reader (text scale 65, background threshold 0.1 and foreground-area filtering at 0.1), excluding query bias from the balancing Gram while retaining the key-bias contribution raises mIoU from 42.411 to 57.527, versus the original reader's 58.010. An SVD implementation of the same bias-inclusive construction reproduces 42.411. Actual Q/K biases remain present and transform with their projections in all arms. Recovery is 15.116 pp [11.854, 19.154]; the recovered-minus-original difference is -0.483 $[-0.855, -0.111]$, leaving noninferiority within 0.5 pp unresolved. These paired-image intervals use 99.1667% coverage in the six-comparison family.

The same-object SVD comparison and the controlled bias-statistic intervention identify the balancing criterion as a substantive design choice. Actual query bias is transported in every arm. The historical SVD and Weight-Gram studies also used different inference pipelines, so their raw cross-study scores do not form a matched method ranking.

C.4.2 Weight-Gram Balancing and Affine CAQC

Weight-Gram balancing uses only the linear projection weights to define the statistic and transports the actual Q/K biases with the projections. The later affine CAQC construction additionally fixes key translation by setting key bias to zero before balancing. For each remaining full-column-rank head, let $S_Q = W_Q^\top W_Q$ and $S_K = W_K^\top W_K$. We use the positive-definite solution $PS_QP = S_K$, set $B = P^{1/2}$, and rewrite $W_Q \leftarrow W_Q B$ and $W_K \leftarrow W_K B^{-\top}$, including the same right action on the query bias while key bias remains zero. Related proposals include QR partial gauge fixing and GaugeKV's Q/K balancing (Wang & Wang, 2026b;c). We derive the displayed map from the SPD equation and use it to obtain affine Q/K consistency of dense self-metrics.

On the frozen 256-image ViT-B/32–ADE pilot, raw dense-SPD gauges change 42.227%, 21.479%, and 22.543% of SCLIP, NACLIP, and PEARL PA+KK pixels. Diagonal RMS balancing still changes 42.668%, 24.667%, and 22.959%. Weight-Gram canonicalization reduces all three below the reported numerical floor, with identity utility changes of only -0.120, -0.084, and -0.025 points.

In the earlier Weight-Gram evaluation on all 1,449 VOC validation images, the canonical identity changes are $+0.263$ points for SCLIP (95% CI $[+0.024, +0.504]$), $+0.138$ for NACLIP ($[+0.028, +0.243]$), and -0.790 for PEARL PA+KK ($[-0.941, -0.628]$). Their raw dense-SPD counterparts lose 6.936, 9.136, and 4.868 points and change 9.425%, 8.915%, and 10.457% of pixels; every canonical gauge comparison changes at most 0.00003% of valid pixels. A pinned NACLIP official-code wrapper gives the same conclusion under the affine audit. Key translation alone changes 0.627% of pixels and loses 0.394 points. Combining it with dense-SPD reparameterization loses 9.279 points and changes 7.675% of pixels, whereas the affine-canonical checkpoint costs only 0.112 points (95% CI $[-0.259, +0.026]$) and leaves 0.0001% numerical disagreement.

A second official end-to-end confirmation keeps ResCLIP's pinned RCS, final self-correlation, SFR, internal PAMR, alias grouping, interpolation, thresholding, and final-mask path intact. On all 1,449 VOC images, the raw combined affine gauge changes 9.099% of pixels and loses 6.557 points (95% CI $[-7.299, -5.753]$). Canonicalization changes identity by $+0.140$ points ($[+0.056, +0.222]$), leaves 0.000048% disagreement, and removes 99.999475% of the raw effect. Native attention, block output, and RCS remain within the specified numerical tolerances.

The offline ViT-B/16 rewrite takes 157 ms and introduces no inference-time module or FLOP. Native error is at most 7.5×10^{-6}. Orthogonal, diagonal, dense-SPD, shear, and nonsymmetric transforms through condition number 17.9 leave canonical dense features within 1.1×10^{-6}. Input-conditioned activation, centered-activation, and spatial-difference Gram variants were stopped after a disjoint 128-image discovery set because their mean utility was 1.307, 1.435, and 0.912 points below the weight construction.

Key-translation audit. A separately frozen 16-image ViT-B/32 audit tests the affine omission directly. The released checkpoint's final-head key-bias RMS is 4.30×10^{-3}. A fixed per-head translation changes native attention probabilities by at most 4.47×10^{-7}, yet moves raw $KK^\top$ by 2.401 and normalized dense features by 0.0247 in maximum absolute value. Zero-key-bias plus Weight-Gram canonicalization reduces both drifts below the recorded numerical floor. Under a combined nonsymmetric $\mathrm{GL}(d)$ gauge and key translation, raw $KK^\top$ and dense feature drifts rise to 71.27 and 0.278, while canonical QQ/KK drifts remain below 4.58×10^{-5} and canonical dense-feature drift is 6.26×10^{-7}. This experiment tests consistency under the affine Q/K subgroup.

C.4.3 An Invariant Intermediate-Feature Consumer

We evaluate a variant of SC-CLIP (Bai et al., 2025) whose attention weights use only intermediate-feature affinity. On 2,000 ADE20K validation images with ViT-B/16, it obtains 20.030 mIoU versus 20.118 for the official SC-CLIP baseline: a difference of -0.087 points, with a paired 95% interval $[-0.173, +0.016]$. Under the audited dense-SPD gauge, disagreement falls from 8.2953% to 0.0003%. This supports a useful invariant consumer with a small observed cost. Its SC-CLIP pipeline, evaluation cohort, and intervention differ from the SCLIP SVD pilot and CAQC experiments, so these scores do not provide a matched ranking. The archived evaluation contains aggregate results and the original interval; per-image arrays are not available in that archive.

C.4.4 Official-Pipeline Calibration and COCO-Stuff Evaluation

The COCO-Stuff evaluation uses all 5,000 `val2017` image/label pairs and OpenAI CLIP ViT-L/14, neither used during method selection. The protocol is fixed before evaluating this panel. NACLIP uses commit `0cac3a651f753b11315ea799cfbfabf79da1bd76`. With PAMR enabled on every image, the unmodified official pipeline reproduces the published ViT-B/16 result of 25.69 mIoU and obtains 24.59 with ViT-L/14.

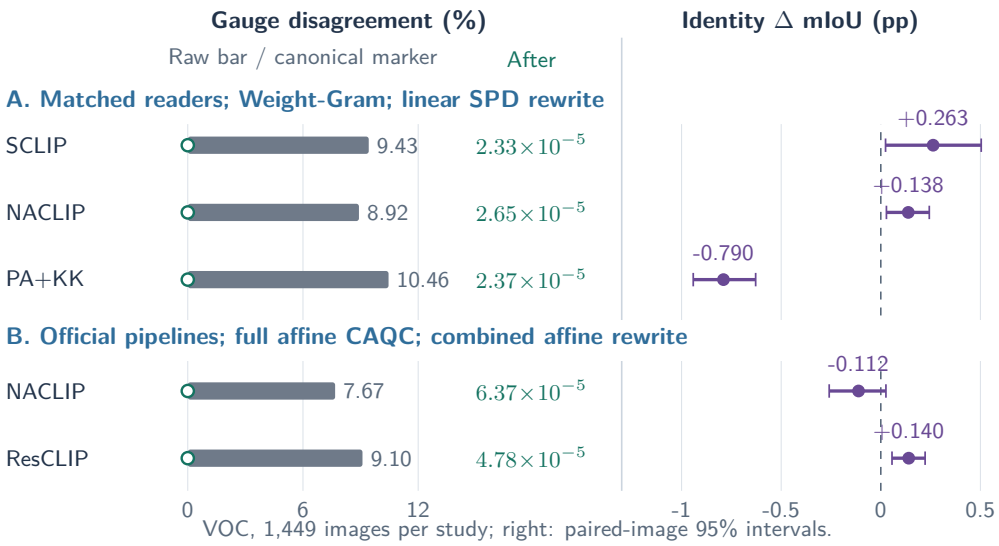

Figure 5: Consistency and task utility on VOC. Left: prediction disagreement under reparameterization, measured against the corresponding raw or canonical identity. Right: canonical-minus-original identity mIoU, with paired-image 95% intervals. Panel A uses the earlier Weight-Gram construction and matched readers; panel B uses complete affine CAQC and official pipelines. Each panel retains its own construction and evaluation protocol.

Table 18: Official NACLIP on 5,000 COCO-Stuff images. Raw/canonical identity mIoU is 24.593/24.660; canonical-minus-raw is $+0.066$ points (95% CI $[+0.035, +0.091]$). Gauge effects below use the identity of the same representation.

Representation	Gauge	Δ mIoU (pp)	Disagreement (%)
Raw	Key translation	$+0.024$	2.555
Raw	Combined affine	-1.377	23.362
Canonical	Key translation	$+0.000$	0.000
Canonical	Combined affine	-0.000007	0.000305

The six conditions use FP32 with the same official data preprocessing, sliding-window inference, PAMR, prompts and evaluator; only the final visual Q/K parameters change. A separate FP16 raw-identity arm gives identical valid-pixel predictions to the unmodified official L/14 implementation. Across the diagnostic subset, maximum native-attention and image-embedding drift are 1.15×10^{-6} and 1.91×10^{-6}; maximum canonical dense-feature drift is 2.08×10^{-6}.

The positive-diagonal construction leaves substantial dense-SPD disagreement, whereas the full-matrix construction reduces it below 0.0002% for all three readers. The full-matrix step therefore removes residual actions left by diagonal balancing.

Table 19: ViT-B/32 evaluation on 1,360 ADE validation images (IDs 641–2000) held out from method selection. Raw identity mIoU is 12.645/15.325/15.431 for SCLIP/NACLIP/PA+KK. Identity deltas compare with raw; gauge deltas compare within each representation. Intervals are 95%, and mIoU differences are pp.

Method	Reader	Identity Δ [interval]	SPD pixel disagree (%)	SPD Δ mIoU
Raw	SCLIP	+0.000	40.751	−0.173
Raw	NACLIP	+0.000	20.426	−1.376
Raw	PA+KK	+0.000	21.169	−1.377
Diag. RMS	SCLIP	+0.280 [+0.190, +0.376]	40.881	−0.560
Diag. RMS	NACLIP	+0.113 [+0.045, +0.175]	22.926	−1.623
Diag. RMS	PA+KK	−0.033 [−0.067, −0.003]	21.650	−1.401
W-Gram	SCLIP	−0.055 [−0.094, −0.019]	0.000130	+0.000
W-Gram	NACLIP	−0.088 [−0.121, −0.054]	0.000118	+0.000
W-Gram	PA+KK	−0.011 [−0.030, +0.006]	0.000082	+0.000

C.5 Source-Scope Intervention

The source-scope diagnostic changes only redistribution among allowed sources for eligible query rows. Each S scope contains reliable same-class sources derived from ground truth; the paired random R scope matches allowed count and coarse distance bins. Contributions of CLS/unknown sources, eligible-source total attention mass, and ineligible rows remain fixed. SCLIP reconstructs each QQ/KK branch before summation. NACLIP scales only its KK term, leaving its spatial prior unchanged. The selected (s_S, s_R) values are $(0.25, 0.5)$ for SCLIP and $(0.125, 0.25)$ for NACLIP; the NACLIP S choice is a grid boundary. The all-query image-equal coverage is 0.800916 and foreground coverage is 0.641896. Each foreground class has eligible queries in at least 15 images. These restrictions do not undo information mixed into V by earlier layers.

Table 20: Label-assisted source-scope diagnostic, rank-16 G. D_S and D_R are TG−T under reliable same-class and matched random allowed sources; $J = D_R - D_S$. Intervals belong to a 12-comparison family (99.5833% each), conditional on three selected seed/bank pairs.

Reader	D_S	D_R	J [interval]
SCLIP	+1.018	+1.907	+0.889 [0.540, 1.215]
NACLIP	+0.782	+1.973	+1.190 [0.975, 1.417]

Temperature is selected separately for each reader/scope, followed by matched training. This is a diagnostic, not a deployable label-free selector.

The G increment is larger in R for both readers (Table 20), with three of three paired effects in the predicted direction. However, S also retains a resolved > 0.5-point benefit: its intervals are $[0.611, 1.422]$ for SCLIP and $[0.600, 0.962]$ for NACLIP. G therefore cannot be characterized solely as removing currently wrong-class sources. Same-class patches can carry mixed earlier-layer information, and changing the source set also changes baseline accuracy and remaining error. These effects prevent a causal percentage attribution to source reliability.

C.6 Earlier Unit-Strength and Joint-LoRA Comparisons

The initial source-trained candidates use three epochs over the same 256 training images, accumulate four image losses, and retain 192 optimizer updates. The standard-LoRA study uses AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, zero weight decay, gradient-norm limit 5, FP32, and TF32 disabled. Its four learning rates are $\{0.0005, 0.002, 0.005, 0.02\}$; its five seeds are 20260821, 20260831, 20260841, 20260851, and 20260861. The smaller temperature and source-scope studies use the first three seeds and rates $\{0.005, 0.02\}$. A candidate tie within 10^{-9} mIoU points is resolved by earlier epoch, then smaller learning rate. Identity remains eligible.

In the PyTorch weight convention, standard LoRA uses $W_Q^L = W_Q + B_Q A_Q$ and $W_K^L = W_K + B_K A_K$, where each full projection is 768×768. Each A is $r \times 768$, each B is $768 \times r$, and Q/K factors are independent. The multiplier α/r is one, dropout is zero, and both Q/K bias increments are trainable. Initialization draws A uniformly in $[-1/\sqrt{768}, 1/\sqrt{768}]$ and sets B and bias increments to zero. This gives 13,824 parameters at rank 4 and 26,112 at rank 8. Combining LoRA4 with G16 gives 26,304, only 192 more than LoRA8. The training budget totals 80 trajectories and 15,360 optimizer steps.

Joint C applies G after its own current LoRA projections. Its preservation reference is that same LoRA state, not an independently optimized L4 model. The comparison tests parameter allocation under the specified training recipe. Five selected endpoints lie at the largest learning rate in the frozen grid. No post-evaluation grid expansion was performed.

Table 21: Matched SCLIP, fixed unit strength, five seeds and equal per-arm search budgets. SD describes seed variation. Preservation refers to original CLIP; joint C preserves only its own current LoRA component under G.

Arm	Parameters	mIoU mean ± SD	Native guarantee
N	0	58.782	Yes
G16	12,480	62.502 ± 0.160	Yes
L4	13,824	61.073 ± 0.567	No
L8	26,112	62.267 ± 0.735	No
C: G16+L4	26,304	61.666 ± 0.581	No

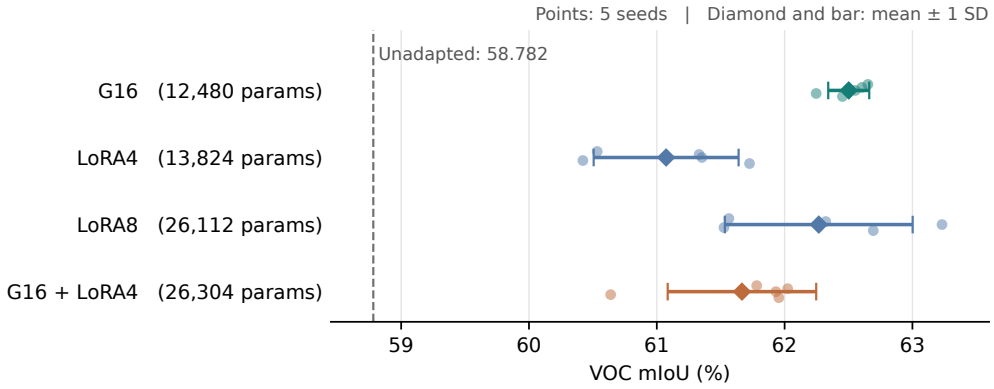

Figure 6: Three-epoch adaptation at unit update strength. Dots are individual seeds; bars are one sample standard deviation. The strength-calibrated comparison in Table 3 uses a six-epoch selection budget.

In the earlier three-epoch unit-strength study, joint G16+LoRA4 scores 61.666 versus LoRA8's 62.267. Their difference is -0.601 $[-1.630, 0.428]$ pp (98.3333% paired-seed interval). Adding G to frozen LoRA8 parents instead gives 0.079 $[-0.477, 0.634]$ pp (97.5% primary interval). These are distinct from the fixed-output-head procedure. The separate 32-training-image comparison at unit strength appears in Section C.10.

Table 22: Protocols for supplementary adaptation comparisons. Image intervals condition on selected states. Seed intervals condition on the evaluation panel. These studies use protocols distinct from the principal strength-calibrated comparisons.

Study	Reader and metric source	Selection	Primary uncertainty
Scalar calibration	Matched SCLIP/NACLIP; host-specific G8	Source dev: scalar, LR, epoch; three seeds	Images; four contrasts, 98.75%
Standard LoRA composition	Matched SCLIP; G16, full-weight LoRA4/8	Source dev: LR, epoch; five seeds; $s = 1$	Seeds; three contrasts, 98.3333%
Source restrictions	Matched SCLIP/NACLIP; G16 and G8; GT-assisted scopes	Source dev: scalar per scope, then LR/epoch; three seed/bank pairs	Images; twelve contrasts, 99.5833%
Cross-reader reuse	Complete ResCLIP/PEARL; frozen NACLIP G16 and G8	Target-reader dev: one scalar per rank, shared across three source seeds	Images; six contrasts, 99.1667%
External reuse	Complete ResCLIP; three frozen NACLIP G8 states	No target fitting; deploy at $s = 1$	Images; twenty-one contrasts, 99.7619%
Frozen LoRA augmentation	Matched SCLIP; selected LoRA8 parent and G16	Two stages, development-only parent and child selection	Five seeds; separate primary and budget-control families

C.7 Adaptation from Selected Parents

The experiment has two stages. Stage one selects a standard LoRA8 parent A and a G16 parent on the same training/development split. We reuse parents from the same model source, training implementation, data order and candidate search. Development data select each parent before the corresponding second stage.

Stage two compares: AG, which freezes A and trains a fresh identity-initialized G16; AC, which continues training LoRA8; AE, which expands A to LoRA12 with zero initial added contribution and then trains all factors; and GG, which continues the same G16 parent rather than adding a second G. All second-stage branches reset the optimizer and use the same prescribed continuation image order, five seeds, four learning rates, and three epochs. AG preserves A's native function; AC and AE may change it.

Table 23: Parameter counts for the second-stage adaptation arms. GG continues its existing G; it does not double its parameter count.

Arm	Total adapter parameters	Stage-two trainable	Preserved reference
A	26,112	0	Frozen selected A
AG	38,592	12,480	A
AC	26,112	26,112	No fixed-function constraint
AE	38,400	38,400	No fixed-function constraint
GG	12,480	12,480	Original CLIP

Stage two uses 80 training trajectories and 15,360 optimizer steps. The arms have equal search and step budgets, although their trainable parameter counts differ. The primary family is A$-$N> 0 and AG$-$A> 0.5 points, using paired five-seed 97.5% Student-t intervals and at least four positive seeds. The budget family compares AG with AC, AE, and GG at > 0.5 points, using 98.3333% intervals and the same sign requirement. These comparisons are confirmatory only if both primary hypotheses

are supported. Mean mIoU is 58.782 (N), 62.267 (A), 62.346 (AG), 63.265 (AC), 63.174 (AE) and 62.907 (GG). The primary AG−A increment is $+0.079$ $[-0.477, 0.634]$ points; A−N is $+3.485$ $[2.337, 4.634]$. The AG−A interval includes zero, leaving the additional benefit of G unresolved. The budget comparisons are therefore secondary; their intervals also include zero for AG−AC and AG−AE.

C.8 Matched Dense-Attention Perturbation

For method m, paired seed j and image i, let $A_{m,j,i}(s)$ stack the dense reader's attention matrices over all crops and heads, including CLS and self entries, and let $A_{N,i}$ denote the original reader's stack. The image-equal average relative perturbation is

$$D_{m,j}(s) = \frac{1}{512} \sum_{i=1}^{512} \frac{\|A_{m,j,i}(s) - A_{N,i}\|_F}{\|A_{N,i}\|_F}.$$

Targets are $d_{\text{high},j} = 0.9 \min\{D_{\text{G16},j}(1), D_{\text{LoRA4},j}(1)\}$ and $d_{\text{low},j} = d_{\text{high},j}/2$. Both methods are calibrated to these targets without labels. Table 24 lists the low- and high-perturbation targets for each paired seed.

Table 24: Locked per-seed relative-attention targets. Targets are dimensionless and chosen without labels.

Paired seed	Low target	High target
20260821	0.285644	0.571288
20260831	0.275831	0.551663
20260841	0.305076	0.610152
20260851	0.302883	0.605767
20260861	0.305435	0.610869

For each of five paired endpoints, we select one coefficient per method at each of two perturbation levels, yielding 20 coefficients. Coefficients are selected without labels on the same 512-image panel subsequently scored, making this a transductive strength-matching diagnostic.

Bounded bisection on $s \in [0, 1]$ finds coefficients without labels; the pair's qualification requires at most 1% relative mean-dose mismatch. It is not a per-row or per-head match. G scales log-eigenvalues, $\lambda \mapsto s\lambda$, with the reciprocal inverse retained; L4 scales both the full additive weight update and its bias increment.

Let I be G's low-to-high mIoU increment minus L4's, and let Q_ℓ, Q_h be the two same-dose G−L4 differences. The three-comparison family uses paired five-seed Student-t intervals with coverage $1 - 0.05/3 = 98.3333\%$ and four degrees of freedom. $I = -0.403$ $[-0.762, -0.043]$ points; $Q_\ell = -1.441$ $[-1.859, -1.022]$ and $Q_h = -1.844$ $[-2.524, -1.163]$. LoRA4 retains an advantage at both matched perturbation levels, with a larger low-to-high increment than G.

Table 25 reports both perturbation levels, with their definition and per-seed targets in Section C.8.

Table 25: Matched average dense-attention perturbation in matched SCLIP with OpenAI CLIP ViT-B/16, evaluated on a fixed panel of 512 VOC evaluation images. Targets are defined separately for each paired seed and shared by G16 and LoRA4; the target column reports their mean and [minimum, maximum] across five pairs. D is a dimensionless relative Frobenius perturbation, defined in Appendix C.8 together with the individual targets. Performance intervals are paired-seed 98.3333% intervals in a three-comparison family. Bold indicates the highest mean within each matched comparison.

Perturbation level	Target D, mean [min, max]	G16 mIoU	LoRA4 mIoU	G16 − LoRA4, pp [interval]
Low	0.294974 [0.275831, 0.305435]	61.234	**62.674**	−1.441 [−1.859, −1.022]
High	0.589948 [0.551663, 0.610869]	62.315	**64.159**	−1.844 [−2.524, −1.163]

C.9 Early-Budget Comparison

The early-budget study reuses recorded training states and observes epochs $1, 2, 3, 6$, with four learning rates per method and four strength choices $s \in \{1/4, 1/2, 3/4, 1\}$. It freezes each choice using 64 development images and evaluates 512 images only afterwards. The common development target is two mIoU points above the original reader, approximately 51.109%. Each method reaches it by epoch two in all five seeds. At the selected checkpoints, G, L4 and N obtain 62.372, 63.746 and 59.323 mIoU, respectively. The 98.75% paired-seed intervals for the four-comparison family give G−L4 $= -1.374$ $[-3.214, 0.466]$ points and G relative cost saving $= 11.33\%$ $[-46.96, 69.63]$; the superiority/noninferiority requirements remain unresolved, and only two seeds have lower G cost. Both hit-minus-native gains exceed one point under the same family.

Costs combine historical training timing across all learning-rate candidates with freshly timed validation, including candidate construction. Shared prefix preparation, model loading, final evaluation, and queueing are outside this estimand. These costs describe the training-and-validation budget of the recorded search.

C.10 Limited-Data LoRA Comparison at Unit Strength

Limited-data adaptation without post-training strength calibration. In a separate 32-image VOC training comparison, G16 obtains 62.005 mIoU versus 59.834 for the full-projection Q/K LoRA4 baseline. The two methods use 12,480 and 13,824 trainable parameters, respectively. The paired improvement is 2.172 points (97.5% per-contrast seed interval [1.133, 3.210]), with all five paired differences positive. Both methods use the same 64-image development set, a fixed 192-update budget per training trajectory, and equal per-method search budgets; endpoints are evaluated at unit update strength, as specified by this protocol, without post-training strength calibration. This establishes a local near-parameter-budget advantage in the stated setting while G preserves the native CLIP function by construction.

The training-label budget is 32 images, whereas the combined training and development label budget is 96. The five repetitions jointly vary training subsets and optimization randomness, and the evaluation panel was used in earlier experiments. The difference between the 32- and 256-image advantages is 0.743 points with interval [−0.689, 2.175], leaving the label-budget trend unresolved. Strength-calibrated comparisons retain their separate protocols and panels in Sections A.5 and C.9.

Table 26: Separate limited-data comparison without post-training strength calibration. Five paired seeds, 32 training and 64 development images, 192 updates per trajectory. G16 minus LoRA4 is +2.1717 pp, with a 97.5% per-contrast paired-seed interval [1.1331, 3.2102] (two-comparison family, four degrees of freedom).

Method	Parameters	Training images	mIoU
G16	12,480	32	62.0052
Full-projection Q/K LoRA4	13,824	32	59.8335

C.11 Joint Fitting with Output Heads

All three arms train from identity for six epochs on train256, with five seeds. The head-only rates are $\{0.0005, 0.002, 0.005, 0.02\}$; the joint arm uses four ordered (G, head) rate pairs: $(0.02, 0.002)$, $(0.08, 0.002)$, $(0.02, 0.005)$ and $(0.08, 0.005)$. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, zero weight decay, constant rates and gradient-norm limit 5. Four-image accumulation gives 64 updates per epoch. The common loss is $0.5\,\mathrm{CE} + 0.5(1 - \mathrm{softIoU})$; the joint arm adds G's $0.001\ \mathrm{mean}(\lambda^2)$ regularizer. Training strengths are one. Hout12 and the joint arm share their initial rank-12 head; Hout24 extends the same seeded factor to rank 24.

Selection observes epochs $\{1, 2, 3, 6\}$. Each head-only endpoint has eight strengths $s_H \in \{0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1\}$; each joint endpoint has eight pairs $(s_G, s_H) \in \{0.5, 1\} \times \{0.25, 0.5, 0.75, 1\}$. Thus each method/seed has $4 \times 4 \times 8 + 1 = 129$ candidates including the unchanged reader, selected by dev64 pooled mIoU. Ties within 10^{-9} points favor the unchanged reader, earlier epoch, lower configuration index, strength closer to one (G first for the joint arm), then stable candidate ID. All states are selected before evaluation. The five Hout12 selections are the parents in Section A.6.

Joint G16+Hout12, Hout12 and Hout24 obtain 67.118, 66.476 and 65.799 mIoU. G16+Hout12 minus the two head-only arms is 0.642 $[-0.995, 2.279]$ and 1.320 $[-1.152, 3.791]$ pp. Intervals have 98.3333% coverage over five paired seeds in a three-comparison family. Both increments remain unresolved.

C.12 Prediction Retention for a Selected Task

The replay uses the selected rank-8 G endpoint (seed 20260821, 6,240 optimization parameters), the OpenAI ViT-B/16 checkpoint, and the matched NACLIP reader at unit multipliers. The 128 VOC images belong to the previously evaluated confirmation partition. The endpoint and five source transformations are fixed before this comparison. Each source includes a shared key translation, including the identity source, which therefore differs from N.

We compare N and the selected target $T_\star$ with five implementations for each source: raw, CAQC-only, naive task reuse, CTGC and full target-Q/K copying. Full-copy implementations use the same target parameters and reader configuration. N, $T_\star$, and the canonical reference obtain 58.791476, 62.046708, and 58.908756 mIoU, respectively. Table 27 reports all sources.

Table 27: Fixed-artifact matched-NACLIP replay on 128 images. Disagreement is pooled over 21,467,621 valid pixels. CTGC is compared with $T_\star$; CAQC-only is compared with CAQC(identity source). The sources are interventions, not optimization seeds or independent replications.

Source	Naive disagreement (%)	Naive$-T_\star$ (pp)	CTGC flips	CAQC flips
Identity	0.000014	−0.000016	0	0
Diagonal, $\kappa = 2$	0.053508	−0.012351	2	4
Dense SPD, $\kappa = 4$	0.074810	−0.029643	2	3
Nonsymmetric shear	0.085967	−0.004777	2	3
Orthogonal	3.614019	−2.651687	0	4

On the prespecified orthogonal source, naive reuse changes 775,844 pixels (3.614%). Across all five sources, CTGC and CAQC-only have maximum continuous probability differences of 2.206×10^{-6} and 1.967×10^{-6}, with elementwise tolerance $2 \times 10^{-5} + 10^{-4}|p_{\text{ref}}|$. Maximum mask disagreement fractions are 9.317×10^{-8} and 1.864×10^{-7}, below the 10^{-5} limit, and maximum absolute mIoU differences are 1.594×10^{-5} and 1.924×10^{-5} points, below 0.01 points. All changed pixels lie within the measured numerical error of a decision boundary. Full target-Q/K copying gives identical scores and masks to $T_\star$.

On the evaluated crops, maximum centered-QK and final-block differences are 1.526×10^{-5}, attention difference is 2.742×10^{-6}, and normalized-CLS difference is 4.173×10^{-7}. On the first eight evaluation images, packed and separately sliced Q/K/V projections produce identical masks.

Payload and scope. The compact runtime tensors use 199,680 bytes; pivots and two hashes add 6,144 and 64 bytes, totaling 205,888 (4.3577% of the full target Q/K tensor payload). The compressed runtime NPZ and descriptive metadata JSON are 187,712 and 34,075 bytes, reported separately. Full Q/K copying stores the selected target's Q/K weights and biases.

For this selected affine task state, transporting the metric and effective bias together retains the target predictions across the five equivalent sources within the stated numerical tolerances.

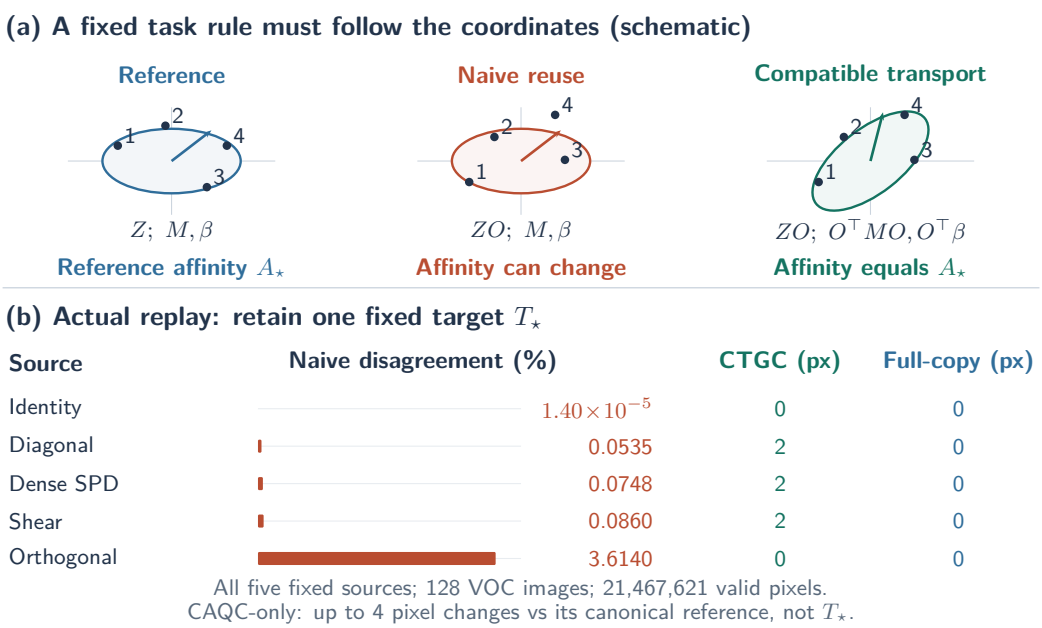

Figure 7: Retaining a selected task across equivalent parameterizations. Top: a schematic illustrates how transporting both the metric and effective bias preserves affinity after a coordinate change. Bottom: actual replay of the fixed endpoint $T_\star$ across all five sources. Naive reuse reports changed-pixel percentages; CTGC and full-copy columns report exact changed-pixel counts. The target is the selected task endpoint, while CAQC-only consistency uses its own canonical reference.

C.13 Illustrations and Qualitative Examples

Figure 8 distinguishes canonical consistency, metric selection and task retention. Figure 9 shows the G parameterization; Figure 1 in the main text gives the PEARL qualitative witness.

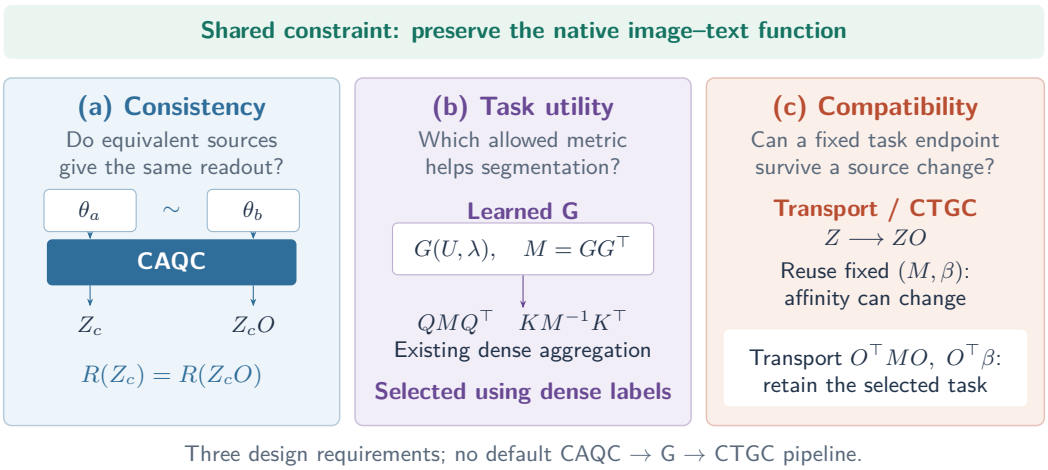

Figure 8: Three questions within a native-function-preserving family. CAQC leaves a shared orthogonal action O that qualifying readers cannot distinguish. Learned G uses dense supervision to select a useful metric. Transport preserves the selected affinity by transforming its metric and effective bias with the coordinates. These are separate constructions; consistency does not select a task-optimal metric, and a selected metric is not automatically compatible with a new source.

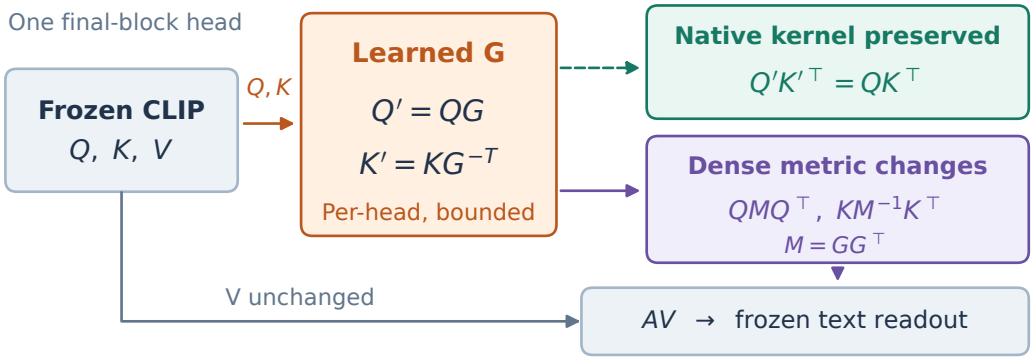

Figure 9: Detailed structure of Learned G. Orange denotes trainable parameters; original CLIP weights are frozen. The green native-QK identity is a preservation property, not another segmentation branch. The selected transform folds into weights and biases.

C.13.1 Qualitative Examples on VOC

Figures 10 and 11 show VOC images 2009_001536 and 2008_007548 (first two in the prespecified image order; seed 20260821). They use the projection comparison's strength-selected G and separately selected scalar/shape controls, respectively. Full-image predictions share VOC colors; light-gray ground truth is ignored.

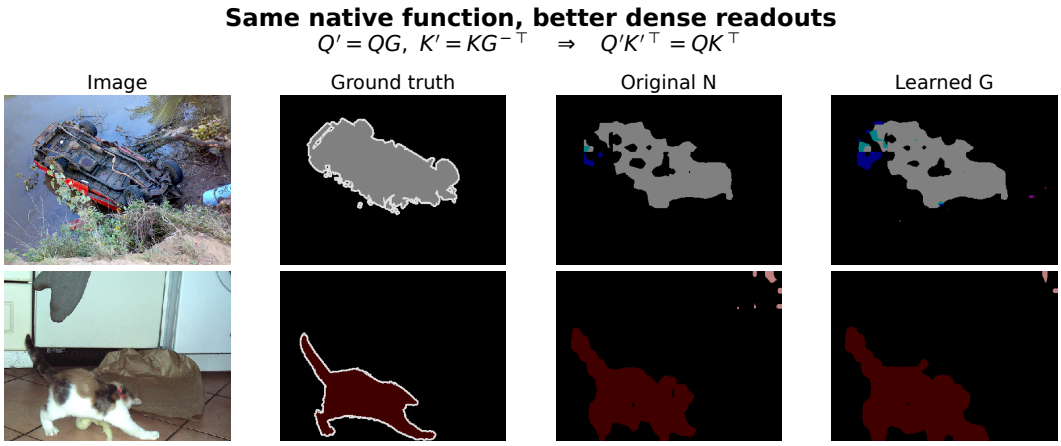

Figure 10: Original-reader and learned-G predictions for the fixed examples. Across the complete 1,321-image VOC evaluation set, the corresponding five-seed means are 58.782 and 62.823 mIoU.

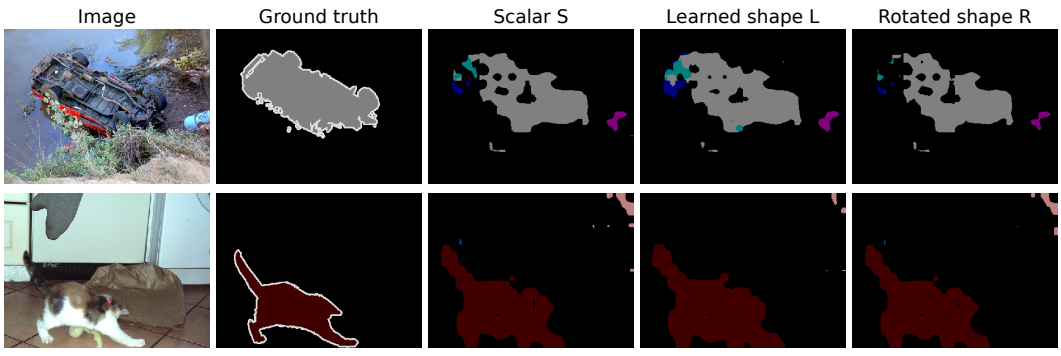

Figure 11: Scalar-only S, learned shape L and rotated shape R on the same examples. L and R inherit their fixed shapes from preceding G training; all three arms receive the same subsequent scalar recalibration. Figure 3 reports the aggregate paired comparisons.

D Implementation, Statistics and Reproduction

D.1 Provenance of the Two G Baselines in the Adapter Comparison

Table 3 retains the G baseline trained and selected within each comparison. The LoRA row reuses earlier G/LoRA trajectories, restricting candidate selection to their first six epochs; later source checkpoints are ineligible. The output-head row trains G anew for six epochs. Both use the same VOC split, five seed values, per-seed image orders, task loss and regularizer, G learning rates $\{0.002, 0.005, 0.02, 0.08\}$, checkpoint epochs $\{1, 2, 3, 6\}$ and strengths $\{0.25, 0.5, 0.75, 1\}$ plus identity.

The implementations compute equivalent projections through different numerical paths: the G/LoRA training code evaluates Q, K and V as separate linear projections, while the output-head study uses a packed QKV projection followed by splitting. The runs therefore have matched nominal hyperparameters but separate G training states and floating-point computation graphs. Their development-selected epochs also differ for one seed (Table 28).

Table 28: Development-selected G configurations for the two rows of Table 3. All selected strengths equal one. Checkpoints with the same nominal configuration still come from their own training runs.

	G in LoRA comparison		G in head comparison	
Seed	Learning rate	Epoch	Learning rate	Epoch
20260821	0.02	6	0.02	6
20260831	0.02	6	0.02	6
20260841	0.08	3	0.08	3
20260851	0.08	6	0.08	3
20260861	0.02	6	0.02	6

D.2 Evaluation Exposure and Statistical Conventions

The VOC evaluation images were used in earlier experiments. The full-panel projection comparison completes the historical 1,321-image evaluation after the earlier 512-image results were known; its interpretation remains post-hoc. The fixed COCO-Object panel was unused at its first ViT-B/16 confirmation and is reused for the later ViT-L/14 evaluation. Neither backbone is fitted or selected on that panel.

Paired-seed intervals describe optimization and selection variability on the fixed evaluation panel. Paired-image intervals condition on selected states. Confidence percentages refer to individual contrasts: Bonferroni families of m comparisons use level $1 - 0.05/m$, targeting nominal 95% family-wise coverage within that family. Class-support conventions are specified for each study.

Joint-study image intervals. The joint-LoRA study uses the $1/120$ and $119/120$ quantiles of 20,000 paired-image bootstrap draws. The resulting intervals are C$-$L8 $[-1.006, -0.204]$, C$-$L4 $[0.376, 0.808]$, and L8$-$N $[2.062, 4.849]$ points. In particular, the C$-$L4 interval includes 0.5 points. These image intervals condition on the selected states; the paired-seed intervals characterize optimization variability.

Post-hoc class-support checks. Historical scorers omit zero-union classes by condition and draw. A separate sensitivity analysis fixes the cohort-GT class set, forms TP/FP/FN from the complete confusion matrix before selecting classes, and assigns zero IoU to empty unions. This sensitivity analysis uses a common class set across conditions; the original scores retain their study-specific class convention. The full-VOC weight-Gram and official NACLIP/ResCLIP cohorts cover all 21 classes and have no audited cross-condition support mismatch. The 1,360-image ADE confirmation covers all 150 classes, but 25 of 10,000 draws differ in support across conditions and 6,910 omit a cohort-GT class. Fixed-support intervals nevertheless round to the displayed values. Official COCO covers all 171 classes with no cross-condition mismatch in its 2,000 draws (seed 20260814). The 256-image weight-Gram pilot covers only 131 ADE classes; its original condition-dependent scores

remain historical selection observations rather than new confirmation evidence. Complete cohort coverage does not imply complete bootstrap coverage.

D.3 Numerical Verification

For the joint-LoRA study, the maximum FP32 relative Frobenius error in $Q_C K_C^\top - Q_L K_L^\top$ is 3.5×10^{-7}, below the 10^{-4} tolerance. An intentionally incorrect inverse produces relative error 0.246. Folded and explicit implementations give the same hard masks on all 84 evaluated endpoints. A separate formula implementation differs at two rejection-boundary pixels, whose threshold distances are of order 10^{-8}. Recomputing confusion matrices from the stored masks reproduces the integer counts exactly.

D.4 Reproduction Materials and Numerical Scope

The accompanying `reproduction.zip` archive covers main Tables 1–4 and Figures 1–3. Its `paper_index.csv` links each entry to the study configuration, selected-state identifiers and hashes, stored sufficient statistics, and a recomputation or reconstruction command. The fixed-head records include the five first-stage parents and their second-stage mappings. From the extracted archive root, `python scripts/recompute.py` rebuilds pooled mIoUs and paired-seed intervals on CPU; `README.md` describes figure reconstruction and dependencies. Recomputation operates on the archived statistics. Benchmark images and pretrained weights are obtained from their official sources rather than redistributed.

The consistency experiments used Python 3.10, PyTorch 2.5.1, torchvision 0.20.1, MMEngine 0.8.4, MMSegmentation 1.1.1, mmcv-lite 2.0.1, and yapf 0.40.1. Nonorthogonal interventions use FP32; upstream-FP16 identity checks are reported separately. Numerical tolerances are stated with the corresponding experiments.