DO SHARED VLM COORDINATES IMPLY SHARED CALIBRATION? A STUDY THROUGH FEW-SHOT CLASS-INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained vision–language models place visual and textual class references in shared coordinates, but alignment does not determine their compatibility with downstream geometric calibration. We expose this distinction through training-free few-shot class-incremental learning, where visual estimates built from labeled images and pretrained textual anchors are evaluated against the same image queries. Crossing visual- and text-derived diagonal calibration transforms with the two isolated scorers reveals a reproducible *role-dependent calibration asymmetry*. Across three encoders and four benchmarks, visual-derived calibration is more favorable to visual than textual scoring in all 12 settings. Textual accuracy decreases in every setting, even when the calibration uses textual statistics from base classes or all currently known classes. Our geometric analysis and source-matched counterexample explain the key distinction: calibration replaces native cosine with a translated, diagonally weighted cosine, and matching reference marginals does not ensure preservation of class ordering for image queries. Together, the empirical asymmetry and geometric analysis motivate source-preserving scoring: calibrate the empirical visual estimate, preserve the pretrained textual anchor, and compose their class evidence. Guided by this principle, we design a simple yet effective **Dual-Space Product-of-Experts (DS-PoE)**, which adds the scores of a base-calibrated visual expert and a native textual expert. **DS-PoE** leads on all four benchmarks and reaches a four-benchmark mean session accuracy of 78.6%, substantially outperforming the previous state-of-the-art training-free method by 3.2 percentage points. Our code is included in the supplementary material.

1 INTRODUCTION

Vision–language models (VLMs) align visual and linguistic representations, supporting multimodal recognition and reasoning across visual question answering, visual grounding, image–text retrieval, and zero-shot recognition (Antol et al., 2015; Goyal et al., 2017; Li et al., 2023; Yu et al., 2018; Kamath et al., 2021; Yan et al., 2024; Wang et al., 2024a; Jia et al., 2021; Radford et al., 2021; Alayrac et al., 2022). Models such as CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), and EVA-CLIP (Sun et al., 2023) encode images and text in shared coordinates. For classification, an image query can be compared with class representations built from example images or text descriptions.

Statistical calibration uses feature statistics to adapt downstream classification scores, with the aim of improving recognition accuracy (Goswami et al., 2023; 2024). Here, geometric calibration centers and rescales features using reference statistics, then normalizes them for cosine scoring. In a VLM, this raises a tempting inference: **because visual and textual class representations share coordinates, the same geometric calibration should be equally compatible with both.** Yet the two representations draw on different evidence: one summarizes observed images, whereas the other derives its relation to images from pretraining. The two reference types may therefore respond differently to calibration; alignment alone does not establish whether both scoring roles benefit. We ask: *do shared coordinates imply shared calibration compatibility?*

We investigate this question through training-free few-shot class-incremental learning (FSCIL). After a data-rich base session, each new class provides only a few labeled examples. These yield a vi-

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.

sual prototype (Snell et al., 2017), while the class description provides a textual reference. Because novel visual prototypes depend on scarce observations, statistics from base classes offer additional structure for calibrating visual scoring without updating the encoders (Goswami et al., 2023; 2024; Wang et al., 2023). This setting also provides a controlled comparison: both reference types represent the same classes and are scored against the same image queries, with encoders held fixed. As classes arrive, the comparison extends to a growing reference set.

To separate the effect of the statistics source from that of the scoring role, we cross a single diagonal centering-and-rescaling family with isolated visual and textual scorers. We estimate calibration from base visual references, base textual references, or all currently known textual references, allowing us to test whether matching the statistics source restores textual discrimination. Each transform acts on both the query and the selected reference, followed by normalization for cosine scoring. We measure compatibility by accuracy changes from native scoring, before score combination.

The study reveals a reproducible **role-dependent calibration asymmetry**. Across CLIP, SigLIP, and EVA-CLIP on four benchmarks, visual-derived calibration is more favorable to visual than textual scoring in all 12 settings, with textual accuracy decreasing throughout. Crucially, **textual scoring deteriorates even under its own statistics**: base-text and session-wise text-derived calibration both reduce accuracy in every setting. Thus, the response is not solely due to statistics from the other modality. **Shared coordinates do not imply shared calibration compatibility.**

To understand why source matching offers no guarantee, we distinguish the statistics of class references from the ordering that recognition requires. Reference statistics describe where class references lie; recognition depends on how an image query ranks them. Centering and rescaling can change this comparison geometry even when applied identically to query and reference. Figure 1 illustrates a source-matched construction in which reference directions remain fixed, but rescaling moves the query across the equal-cosine decision boundary. **Matching reference statistics therefore does not ensure preservation of query–reference class ordering.** Moving from this constructed example to real CLIP queries, we further examine the true class's ordering against all competing classes. For textual scoring, calibration causes a net loss of pairwise wins, linking the geometric issue to recognition.

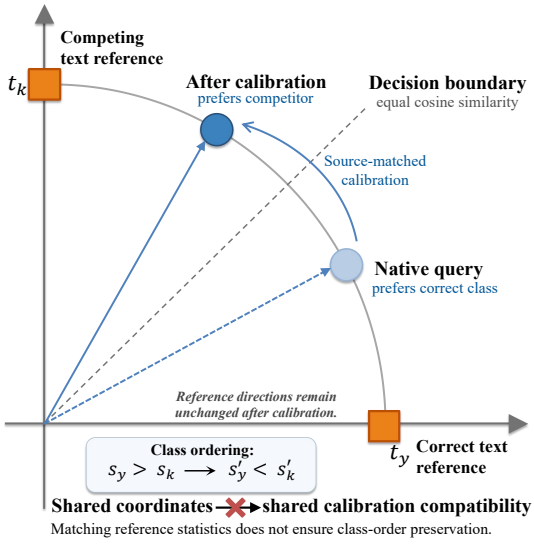


Figure 1: **Source-matched calibration can reverse class ordering.** Schematic of the construction in Appendix A. Query and references share reference-derived rescaling and normalization. Reference directions stay fixed, but the query crosses the equal-cosine boundary, switching from the correct class to its competitor.

Together, the empirical asymmetry and geometric analysis motivate calibrating the support-derived visual estimate while retaining the textual anchor in its native geometry. The two scorers nevertheless produce scores for the same classes, allowing their evidence to be combined without mixing their reference geometries. We instantiate this principle as **Dual-Space Product-of-Experts (DS-PoE)**, a classifier combining scores from a base-calibrated visual expert and a native textual expert. **DS-PoE** leads on all four benchmarks and attains the strongest mean across benchmarks. Composition controls examine the benefit of preserving separate scoring geometries, while support resampling assesses robustness.

Our contributions are ***(i)*** a reproducible role-dependent calibration asymmetry: crossed interventions across three encoders and four benchmarks reveal textual degradation under all three tested transforms and a more favorable visual response to visual-derived calibration in all 12 settings; ***(ii)*** a task-independent geometric account of why source matching is insufficient for cosine-based query–reference scoring: we derive the calibrated cosine geometry and construct a source-matched class-ordering reversal; and ***(iii)*** a simple source-preserving classifier, **DS-PoE**, combining cali-

brated visual and native textual scoring by fixed addition. It achieves state-of-the-art performance on all four benchmarks, with a mean of 78.6% across the four benchmarks, 3.2 percentage points above the strongest prior training-free method, without gradient updates or replay.

2 Related Work

Geometry of vision–language representations. Vision–language models place images and text in comparable embedding coordinates, while the two modalities can retain distinct geometric structure. Liang et al. (2022) characterize the modality gap in contrastive multimodal representations. This distinction is also relevant to training-free adaptation, where statistically constructed visual classifiers have been combined with native textual classifiers (Wang et al., 2024b). Subsequent work studies the gap's structure, origins, and relation to cross-modal alignment (Eslami & de Melo, 2025; Schrodi et al., 2025; Yamaguchi et al., 2025). Recent geometric analyses further reveal modality-specific organization within shared VLM spaces: Mistretta et al. (2025) study differences between intra- and inter-modal similarity structure, while Levi & Gilboa (2025) characterize CLIP embeddings through modality-specific non-centered geometries. These works establish that coordinate comparability need not imply geometric equivalence. More recent methods also explore cross-modal prototype alignment and mapping to reduce modality discrepancy in shared embedding spaces (Goswami et al., 2026; Yang et al., 2026). However, existing alignment and adaptation approaches primarily reduce modality discrepancy or improve recognition, rather than asking whether a common post-hoc calibration family remains compatible with different reference roles. We instead ask whether calibration preserves class ordering and is equally compatible with visual and textual scoring.

Few-shot class-incremental learning and training-free calibration. Few-shot class-incremental learning (FSCIL) requires recognition of an expanding class set when each newly introduced class provides only a few labeled examples (Tao et al., 2020). Existing methods adapt classifiers or representations through continual optimization (Zhang et al., 2021; Zhou et al., 2022; Song et al., 2023; Yang et al., 2023; Oh et al., 2024), and recent VLM-based approaches further employ prompt learning or multimodal adaptation (Park et al., 2024; Sun et al., 2024). We focus on the training-free setting, where frozen VLMs provide visual and textual references for analyzing scoring geometry.

Several training-free FSCIL methods modify class references or their scoring geometry. TEEN (Wang et al., 2023) calibrates novel visual prototypes using semantically related base prototypes; FDR (Xue et al., 2025) uses LLM-generated descriptions and attributes to decompose and recompose visual features; BiMC (Chen et al., 2025) calibrates visual and textual classifiers before constructing mixed category centers; and HyCal (Lee et al., 2026) combines cosine and Mahalanobis geometry for visual prototype calibration under cross-discipline FSCIL. Before combining or calibrating different references, we ask whether the same geometric calibration is equally compatible with their distinct roles. The resulting role-dependent calibration asymmetry motivates **DS-PoE** to calibrate support-derived visual evidence while preserving native image–text scoring.

3 Class References and Role-Dependent Geometric Calibration

We test whether shared coordinates imply shared calibration compatibility through controlled comparisons of visual and textual scoring. The crossed study establishes the direction of their responses; the geometric derivation explains why estimating a calibration transform from the reference source does not enforce preservation of the recognition decision.

3.1 A controlled scoring problem

Let $\mathcal{C}^0, \ldots, \mathcal{C}^T$ be disjoint class sets introduced in successive sessions. The base session provides a labeled training set; each incremental session supplies K images per new class. At session t, evaluation covers $\mathcal{C}_{\le t} = \bigcup_{j=0}^{t} \mathcal{C}^j$. Let f_I and f_T denote the image and text encoders with D-dimensional outputs, and let $\mathrm{Norm}(z) = z/\|z\|_2$. Let $q = \mathrm{Norm}(f_I(x))$ be an image query, v_c a normalized visual prototype, and t_c a normalized textual prototype. Their native scorers are

$$s_c^{V,I}(q) = q^\top v_c, \qquad s_c^{T,I}(q) = q^\top t_c, \tag{1}$$

where I denotes native geometry; only the reference role changes, not the image query.

3.2 CROSSING STATISTICS SOURCES WITH REFERENCE ROLES

We use a reference bank to denote a set of class reference vectors used by one scorer. For a reference bank $\mathcal{R} = \{r_i\}_{i=1}^n$, we estimate a diagonal calibration transform:

$$\mu_{\mathcal{R}} = \frac{1}{n}\sum_i r_i, \qquad u_{\mathcal{R}} = \frac{1}{n-1}\sum_i (r_i - \mu_{\mathcal{R}})^{\odot 2}, \tag{2}$$

$$\widetilde{u}_{\mathcal{R}} = (1-\rho)u_{\mathcal{R}} + \rho\left(\frac{1}{D}\sum_{j=1}^{D} u_{\mathcal{R},j}\right)\mathbf{1}, \qquad W_{\mathcal{R}}(z) = \mathrm{Norm}\Big((z-\mu_{\mathcal{R}}) \oslash \sqrt{\widetilde{u}_{\mathcal{R}}}\Big). \tag{3}$$

Here $\odot$ and $\oslash$ denote elementwise operations. The coefficient ρ shrinks each coordinate variance toward the mean variance, controlling how much dimension-specific rescaling remains. At $\rho = 0$, the transform keeps the full diagonal rescaling; as ρ approaches one, the rescaling becomes isotropic and normalization removes its common scale. The transform is estimated from between-class prototype statistics rather than within-class image covariance.

We form W_V from base visual prototypes, W_{T_b} from base textual prototypes, and W_{T_t} from all textual prototypes of currently known classes. Crossing each statistics source with both reference roles separates source matching from role-dependent calibration response. For $R \in \{V, T\}$ and $S \in \{V, T_b, T_t\}$, the crossed scorer is

$$s_c^{R,S}(q) = W_S(q)^\top W_S(r_c^R), \quad r_c^V = v_c, \quad r_c^T = t_c. \tag{4}$$

Both arguments receive the same transform, so each crossed scorer evaluates query–reference similarity in a common calibrated geometry. W_{T_t} is re-estimated from currently known class texts at each session; the other two transforms remain fixed after the base session. All interventions use the same frozen encoder and isolated scoring rule.

We measure the response relative to each native scorer, writing $A_{\mathrm{avg}}(s)$ for the mean session accuracy of scorer s under the protocol of Section 5.1:

$$\Delta_{R\leftarrow S} = A_{\mathrm{avg}}(s^{R,S}) - A_{\mathrm{avg}}(s^{R,I}), \qquad \Gamma = \Delta_{V\leftarrow V} - \Delta_{T\leftarrow V}. \tag{5}$$

Thus positive Γ means that the visual-derived calibration is more compatible with visual scoring, measured by accuracy change, regardless of whether visual accuracy itself increases.

3.3 ROLE-DEPENDENT CALIBRATION ASYMMETRY

Under the evaluation protocol of Section 5.1, the crossed intervention reveals a consistent role-dependent response across three VLM encoders and four benchmarks. Visual-derived calibration is more compatible with visual than textual scoring in all 12 encoder–dataset settings. Textual scoring deteriorates under the same calibration throughout, and remains degraded when the calibration is instead estimated from either base textual references or all textual references available at the current session. Thus source matching is insufficient to recover native textual scoring.

The same asymmetry appears at the decision level. On CLIP, visual-derived calibration produces a positive pairwise ordering response for the visual scorer on all four benchmarks, whereas all three transforms produce negative ordering responses for the textual scorer. Section 5 quantifies these effects; we next examine why source matching does not guarantee preservation of class ordering.

3.4 WHY SOURCE MATCHING NEED NOT PRESERVE CLASS ORDERING

Calibration changes the comparison geometry. Write $D_S = \mathrm{diag}(\widetilde{u}_S^{-1/2})$. Since D_S is diagonal, $D_S^\top = D_S$ and $M_S = D_S^\top D_S = D_S^2$. Direct substitution into Eq. (4) gives

$$s_c^{R,S}(q) = \frac{(q-\mu_S)^\top M_S (r_c^R - \mu_S)}{\sqrt{(q-\mu_S)^\top M_S (q-\mu_S)}\sqrt{(r_c^R-\mu_S)^\top M_S (r_c^R-\mu_S)}}. \tag{6}$$

This is cosine in a translated, diagonally reweighted coordinate system, or an affine Mahalanobis-like cosine. Before normalization, coordinates with smaller estimated variance are amplified more strongly, changing their relative contributions to the comparison. Applying the same transform to both arguments therefore defines a common calibrated geometry, but need not preserve their native angular ordering.

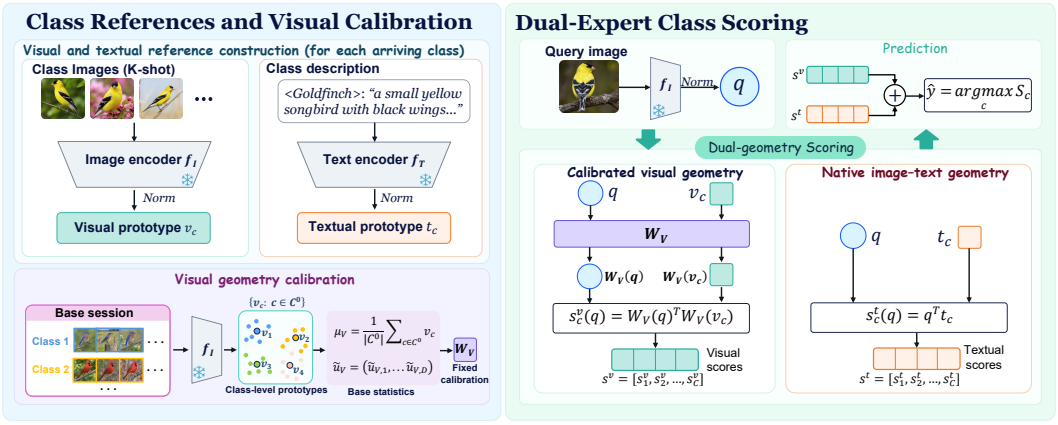


Figure 2: **Source-preserving DS-PoE. Left:** Class images and fixed descriptions form visual and textual reference banks. Visual references are normalized class means built from base training images or K incremental supports; base visual prototypes determine the fixed calibration W_V. **Right:** A shared image query is scored in calibrated visual and native image–text geometries, and the two vectors of class scores are added for prediction.

Relational class ordering. For two class references r_a, r_b, the query normalization is common and positive, so the sign of their calibrated score difference is the sign of

$$(q-\mu_S)^\top M_S\left(\frac{r_a-\mu_S}{\|r_a-\mu_S\|_{M_S}}-\frac{r_b-\mu_S}{\|r_b-\mu_S\|_{M_S}}\right), \qquad \|z\|_{M_S}=\sqrt{z^\top M_S z}. \tag{7}$$

The marginal means and variances in Eq. (2) do not enforce agreement between this sign and that of $q^\top(r_a-r_b)$. Even an exactly source-matched calibration can reverse a class preference: Appendix A gives the exact construction behind Figure 1. In that example, the reference directions remain fixed after normalization, but anisotropic rescaling moves the query across their equal-cosine boundary. Matching marginal reference statistics does not ensure preservation of class ordering, even when the same calibration acts on both arguments of the scorer. Changing ordering is not itself undesirable: visual calibration may improve recognition by doing so. The key is that source matching alone does not determine whether a change benefits a given scorer role.

The crossed experiments establish how calibration changes decisions across reference roles. An empirical visual reference summarizes observations in the target domain, whereas a textual reference is used through its pretrained relation to image queries. Re-estimating a textual marginal leaves that image–text ordering unconstrained. Together, the geometric distinction and the observed role-dependent response motivate source-preserving scoring: calibrate the empirical visual estimate while preserving the pretrained textual anchor.

4 Source-Preserving DS-PoE

DS-PoE implements the source-preserving principle by pairing a calibrated visual expert with a native textual expert. For each currently known class, we store one visual reference and one textual reference; the collection of visual references and the collection of textual references form the two reference banks. Figure 2 separates their construction and the base calibration statistics from per-query prediction. Class references are prepared when their classes arrive; each subsequent query is encoded once and scored against both banks.

4.1 Class References and Visual Calibration

For class c, let $\mathcal{S}_c$ contain its available training images and let d_c contain its class name and fixed description. The image and text encoders construct two complementary references:

$$v_c=\mathrm{Norm}\left(\frac{1}{|\mathcal{S}_c|}\sum_{x_i\in\mathcal{S}_c} f_I(x_i)\right), \qquad t_c=\mathrm{Norm}(f_T(d_c)). \tag{8}$$

The visual prototype summarizes observed class appearance, while the textual prototype provides a pretrained semantic anchor. Base prototypes use all base training images; incremental prototypes use the K support images provided at arrival. In the primary CLIP implementation, image embeddings are averaged before prototype normalization. Each description is encoded once and cached.

We estimate the visual calibration transform from the base reference bank $\{v_c : c \in \mathcal{C}^0\}$ using Eqs. (2)–(3). Its mean μ_V and diagonal variance $\widetilde{u}_V$ describe variation across base class prototypes. The resulting transform is

$$W_V(z) = \mathrm{Norm}\Big((z - \mu_V) \oslash \sqrt{\widetilde{u}_V}\Big). \tag{9}$$

Centering and rescaling define visual scoring geometry; normalization makes scoring a cosine comparison. The transform remains fixed after the base session. We cache $W_V(v_c)$ and t_c in the visual and textual banks, respectively, retaining native image–text geometry for text.

4.2 Dual-expert class scoring

For a test image x, we compute the normalized query $q = \mathrm{Norm}(f_I(x))$ once. The visual branch applies W_V to this query and compares it with the calibrated visual references; the textual branch compares the native query directly with the textual references:

$$s_c^v(q) = W_V(q)^\top W_V(v_c), \qquad s_c^t(q) = q^\top t_c. \tag{10}$$

Thus the same query supports two role-specific comparisons. The visual expert uses base-class statistics to calibrate support-derived evidence, while the textual expert preserves the pretrained image–text comparison. Each branch produces a score vector over the same currently known classes. Because corresponding entries index the same class identities, these score vectors remain directly composable despite being formed in different geometries.

DS-PoE combines these vectors by fixed addition:

$$s_c(q) = s_c^v(q) + s_c^t(q), \qquad \hat{y} = \arg\max_{c \in \mathcal{C}_{\le t}} s_c(q). \tag{11}$$

The references retain their geometries until evidence is combined at the level of class scores. Writing the experts' unnormalized potentials as $\exp(s_c^v)$ and $\exp(s_c^t)$ gives $\exp(s_c) = \exp(s_c^v)\exp(s_c^t)$, the product-of-experts interpretation of fixed score addition (Hinton, 2002). Appendix B gives the normalized identity. Composition uses no learned modality weight or weight chosen for each dataset.

Incremental updates and cost. Each arriving class appends its calibrated visual prototype and native textual anchor to the corresponding banks without changing the encoders, calibration statistics, or existing references. For C known classes, the two banks and the calibration's mean and scale require $2CD + 2D$ scalars, excluding the encoder parameters. After image encoding, prediction requires an $O(D)$ query calibration and scoring against two $C \times D$ banks. No gradient updates, replay, or cache of individual support features are needed. Appendix B gives the complete inference procedure.

5 Experiments

The experiments connect geometry to recognition: we establish cross-encoder calibration asymmetry, examine its effect on class ordering, and evaluate the resulting source-preserving classifier.

5.1 Experimental setup

Benchmarks and metrics. We follow the few-shot class-incremental data schedules of FDR (Xue et al., 2025) on CIFAR-100 (Krizhevsky, 2009), miniImageNet (Vinyals et al., 2016), CUB-200 (Wah et al., 2011), and IN1K-FSCIL (Russakovsky et al., 2015). Their respective base-class counts are 60, 60, 100, and 800; incremental sessions add 5, 5, 10, and 20 classes until all 100, 100, 200, and 1,000 classes are known. Base sessions use all available training images and incremental sessions use five images per class. If A_t is accuracy over all known classes at session t, we report $A_{\mathrm{base}} = A_0$ and $A_{\mathrm{avg}} = (T+1)^{-1} \sum_{t=0}^{T} A_t$. Benchmark means weight datasets equally.

Table 1: Crossed calibration responses: $\Delta_{R \leftarrow S}$ in A_{avg} percentage points relative to each native scorer. V supplies base visual statistics; T_b and T_t supply base and currently known textual statistics. The image query and the selected class reference receive the same transform in each cell.

Encoder	Dataset	$V \leftarrow V$	$T \leftarrow V$	$V \leftarrow T_b$	$T \leftarrow T_b$	$V \leftarrow T_t$	$T \leftarrow T_t$
CLIP	CIFAR-100	+4.10	-2.20	-4.09	-6.55	-3.83	-6.23
CLIP	miniImageNet	+3.07	-16.56	-5.81	-6.58	-5.62	-5.11
CLIP	CUB-200	+1.16	-11.16	-7.38	-51.80	-7.14	-51.09
CLIP	IN1K-FSCIL	+4.19	-21.85	-5.81	-13.31	-5.82	-13.48
SigLIP	CIFAR-100	+2.47	-0.40	-0.71	-51.38	-0.76	-53.63
SigLIP	miniImageNet	+1.20	-15.34	-0.46	-56.62	-0.35	-51.88
SigLIP	CUB-200	-0.15	-14.88	-2.04	-51.53	-1.90	-49.55
SigLIP	IN1K-FSCIL	+1.65	-22.80	-0.23	-67.36	-0.24	-67.33
EVA-CLIP	CIFAR-100	+0.39	-1.42	+0.02	-19.13	-0.02	-18.31
EVA-CLIP	miniImageNet	+0.13	-2.12	-0.04	-12.68	-0.04	-13.44
EVA-CLIP	CUB-200	-0.24	-5.89	+0.07	-16.29	+0.02	-15.04
EVA-CLIP	IN1K-FSCIL	+0.06	-2.67	+0.04	-23.37	+0.05	-23.44
Macro mean		+1.50	-9.77	-2.20	-31.38	-2.14	-30.71

Encoders and controls. The main comparison uses CLIP ViT-B/16. The crossed study also evaluates SigLIP ViT-B/16 and EVA02-CLIP-B/16 with their native preprocessing and tokenization. Following the offline description protocols of FDR (Xue et al., 2025) and BiMC (Chen et al., 2025), class descriptions are generated with an LLM. All scoring variants within a setting use the same descriptions, class references, and test queries. Appendix C details implementation and text controls.

5.2 Crossed calibration reveals role-dependent asymmetry

Visual-derived calibration reveals role asymmetry. Table 1 tests whether a calibration transform estimated from visual references affects both reference roles equally. Across three encoders, W_V is more favorable to visual than textual scoring in all 12 encoder–dataset settings. Its macro response is $+1.5$ points on visual scoring and -9.8 points on textual scoring. The visual response is positive on all four CLIP settings and changes little for EVA-CLIP, whereas the textual response remains negative across all three encoders. This consistent response gap establishes role-dependent calibration asymmetry.

Source matching does not restore textual scoring. Replacing visual statistics with base textual statistics yields a negative textual response in every setting, with a macro change of -31.4 points. Updating those statistics from all currently available class texts likewise degrades textual scoring in all 12 settings (-30.7 points on average). The latter control includes each current textual reference in its own statistics bank. Neither base nor current textual statistics recover native textual scoring, motivating preservation of native text scoring. Controls for the combined classifier appear in Appendix G.

The asymmetry is statistically consistent. We evaluate Γ_{final} with a paired, class-balanced bootstrap on the final-session predictions (Efron, 1979). All 12 final-session interaction estimates are positive, with 95% confidence intervals excluding zero in 11 settings. Appendix E reports the complete point estimates and confidence intervals.

5.3 Class ordering reveals the effect on decisions

Beyond top-1 accuracy, we examine how calibration changes the true class's ordering against all competing classes, using the pairwise comparisons in Eq. (7). From saved CLIP logits, we define the true-class pairwise win rate: for a query with label y and C known classes, $B(s, y) = (C - 1)^{-1} \sum_{c \neq y} \mathbf{1}[s_y > s_c]$. This measure captures changes in the true class's relative position even when the top-1 prediction remains unchanged. Our ordering response Ω is the change in B, averaged first across queries within a session and then equally across sessions. Table 2 shows negative textual

Table 2: Response in CLIP class ordering Ω (percentage points), averaged equally across sessions. Positive values indicate net gains in pairwise wins for the true class; negative values indicate net losses.

Dataset	$V \leftarrow V$	$T \leftarrow V$	$T \leftarrow T_b$	$T \leftarrow T_t$
CIFAR-100	+0.98	-0.19	-1.77	-1.84
miniImageNet	+0.29	-1.23	-0.27	-0.18
CUB-200	+0.02	-1.57	-2.63	-2.73
IN1K-FSCIL	+0.21	-0.81	-0.26	-0.26

Table 3: Training-free few-shot class-incremental learning. Accuracy is in percent. Bold marks the best displayed value. Mean averages the four A_{avg} entries.

Method	CIFAR-100		miniImageNet		CUB-200		IN1K-FSCIL		Mean
	A_{base}	A_{avg}	A_{base}	A_{avg}	A_{base}	A_{avg}	A_{base}	A_{avg}	A_{avg}
TEEN (Wang et al., 2023)	75.7	68.4	92.0	88.7	80.9	73.7	69.2	66.1	74.2
FeCAM (Goswami et al., 2023)	75.7	67.9	91.0	88.0	80.8	73.6	69.8	66.4	74.0
BiMC (Chen et al., 2025)	77.3	72.5	93.8	91.5	75.5	70.9	68.2	66.7	75.4
FDR (Xue et al., 2025)	75.7	68.5	92.0	89.1	80.9	74.1	69.2	66.4	74.5
DS-PoE	**79.2**	**73.1**	**94.3**	**92.4**	**82.4**	**76.0**	**75.5**	**72.7**	**78.6**

ordering responses under all three evaluated calibration transforms on every dataset, indicating a net loss of pairwise wins for the true class. Within this CLIP analysis, the visual-derived transform has a positive ordering response on all four datasets, whereas controls using textual statistics are negative. Appendix F specifies tie handling and aggregation.

5.4 Training-free incremental recognition

Under the training-free protocol in Table 3, **DS-PoE** attains the highest A_{avg} on all four benchmarks, with a mean of 78.6% across the four benchmarks versus 75.4% for BiMC. Its gains over the strongest prior entries are 0.6, 0.9, 1.9, and 6.0 points on CIFAR-100, miniImageNet, CUB-200, and IN1K-FSCIL, respectively. **DS-PoE** also attains the highest A_{base} on every benchmark. The largest gains occur on CUB-200 and IN1K-FSCIL, without encoder updates or replay. We next examine how calibration and score composition contribute to these improvements.

5.5 Why preserve two scoring geometries?

Table 4 separates the roles of geometric calibration and composition. Calibrating the visual expert raises its mean A_{avg} across the four benchmarks from 73.9% to 77.0%; the same calibration reduces the textual expert from 69.2% to 56.2%. **DS-PoE** reaches 78.6%, versus 76.7%, 76.6%, and 77.2% for native-score addition, shared visual calibration, and equal-weight prototype mixing, respectively, and wins on every benchmark (mean gains of 1.9, 2.0, and 1.4 points). Both the calibrated visual estimate and the native textual anchor contribute to the resulting classifier.

Equal-weight prototype mixing evaluates $\text{Norm}(0.5v_c + 0.5t_c)$ with native cosine, providing a direct control for combining the references before scoring. Appendices J and K report sensitivity to prototype mixing and score weighting. **DS-PoE** obtains the highest mean across the four benchmarks among the composition controls in Table 4. These controls distinguish how calibration affects each scorer from how their score vectors are combined. Across all 12 settings, **DS-PoE** improves the macro mean over native-score addition by 1.2 points and over shared visual calibration by 1.1 points (Appendix D).

Complementary class evidence. Of all CLIP queries, 4.8–9.7% are correctly recognized only by the native textual expert and 6.3–24.4% only by the calibrated visual expert, averaging equally across sessions (Table 7). Their usefulness is therefore not determined by comparing standalone accuracies alone: each preserves correct decisions absent from the other.

Table 4: Controlled CLIP ViT-B/16 ablation: A_{avg} (%). All variants within a dataset use identical class references and queries. Bold denotes the best displayed value in each column.

Scoring rule	CIFAR	miniIN	CUB	IN1K	Mean
Visual, native	67.8	88.0	73.5	66.3	73.9
Visual, calibrated	71.9	91.1	74.6	70.4	77.0
Text, native	63.0	89.5	58.0	66.1	69.2
Text, visually calibrated	60.9	73.0	46.7	44.3	56.2
Native-score addition	71.1	90.7	73.7	71.3	76.7
Shared visual calibration	72.5	90.5	72.8	70.5	76.6
Equal-weight prototype mixing	72.2	91.0	74.3	71.4	77.2
DS-PoE	**73.1**	**92.4**	**76.0**	**72.7**	**78.6**

Support variation. Three-seed evaluations retain positive visual calibration responses on all four CLIP datasets. Across the three support seeds, **DS-PoE** obtains mean A_{avg} of 73.1%, 92.1%, 75.8%, and 72.8% on CIFAR-100, miniImageNet, CUB-200, and IN1K-FSCIL, respectively. The 1/2/5/10-shot study further tests how the visual estimate and its combination with fixed text evolve as support increases. Tables 11 and 12 report all available resampling and shot results.

Calibration strength. We use the same shrinkage coefficient $\rho = 0.1$ on every dataset and encoder. Across the four CLIP benchmarks, the three-seed **DS-PoE** means vary by at most 0.13 points for $\rho \in \{0, 0.05, 0.1, 0.2\}$; throughout this range, visual calibration improves the visual scorer and reduces textual accuracy. Appendix I reports the complete strength sweep, including stronger shrinkage and its effect on the role-dependent response.

6 Discussion and Limitations

Our results separate two questions often conflated in VLM recognition: how class references are distributed and how image queries rank them. The geometric argument concerns the query, references, and calibration transform; it does not depend on class-arrival schedules or support size. In training-free few-shot class-incremental learning, this distinction motivates calibrating support-derived visual estimates while retaining pretrained textual anchors in their native geometry. It also suggests assessing post-hoc geometric adaptation by how it changes the ranking of class references for each scoring role, not by reference statistics alone.

Our empirical study examines diagonal centering and rescaling with fixed pretrained encoders. The main calibration setting yields a consistent relative visual–text compatibility gap across all 12 encoder–dataset settings, while response magnitudes and directions can vary with the backbone and calibration strength. Extending these findings to richer post-hoc transforms remains open.

7 Conclusion

Shared embedding coordinates do not imply shared calibration compatibility. Our geometric derivation and source-matched counterexample show that matching reference statistics under a post-hoc transform does not guarantee preservation of class ordering for image queries. Crossed interventions across three pretrained VLM encoders and four benchmarks reveal a consistent role-dependent response: visual-derived calibration is more favorable to visual than textual scoring in all 12 encoder–dataset settings, while textual scoring deteriorates even under text-derived calibration.

DS-PoE turns this finding into a practical source-preserving classifier combining a calibrated visual expert with the native image–text expert. It achieves state-of-the-art performance on all four benchmarks, raising the strongest prior mean across the four benchmarks from 75.4% to 78.6% without gradient updates or replay. Beyond incremental learning, this analysis suggests assessing post-hoc transforms by how they change class ordering for each scoring role, rather than by matching reference statistics alone.

AI USE STATEMENT

Generative AI tools assisted with research discussion, experiment-code drafting and debugging, and manuscript drafting and revision. Class descriptions were also generated offline with a language model as described in Section 5.1. The authors are responsible for the final manuscript, code, and empirical claims.

ETHICS STATEMENT

This work evaluates image recognition on established public benchmarks using pretrained vision–language encoders. It involves no human participants, personal-data collection, user study, or deployment in a consequential decision-making setting. The method operates on class-level image and text references rather than individual-level attributes. As with other systems built from pretrained encoders and benchmark labels, its behavior reflects the semantic coverage and biases of those underlying resources; benchmark accuracy should therefore be interpreted as a recognition measure rather than a fairness or deployment guarantee.

REPRODUCIBILITY STATEMENT

Sections 3–5 specify the crossed calibration analysis, **DS-PoE** scoring rule, benchmark schedules, encoders, and evaluation controls. Appendix A gives the geometric derivation and source-matched counterexample, Appendix B provides the complete inference procedure, and Appendix C records implementation conventions and data controls. The complete implementation and evaluation code are included in the supplementary material.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Mariana Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433, 2015.

Yiyang Chen, Tianyu Ding, Lei Wang, Jing Huo, Yang Gao, and Wenbin Li. Enhancing few-shot class-incremental learning via training-free bi-level modality calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9881–9890, 2025.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in CLIP. In *International Conference on Learning Representations*, 2025.

Dipam Goswami, Yuyang Liu, Bartłomiej Twardowski, and Joost van de Weijer. FeCAM: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. In *Advances in Neural Information Processing Systems*, pp. 15548–15562, 2023.

Dipam Goswami, Bartłomiej Twardowski, and Joost van de Weijer. Calibrating higher-order statistics for few-shot class-incremental learning with pre-trained vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4075–4084, 2024.

Dipam Goswami, Simone Magistri, Gido M. van de Ven, Bartłomiej Twardowski, Andrew D. Bagdanov, Tinne Tuytelaars, and Joost van de Weijer. Cross-modal prototype alignment and mixing for training-free few-shot classification. *arXiv preprint arXiv:2603.24528*, 2026.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, 2017.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. doi: 10.1162/089976602760128018.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 4904–4916, 2021.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr: Modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1780–1790, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Eunju Lee, MiHyeon Kim, JuneHyoung Kwon, Yoonji Lee, JiHyun Kim, Soojin Jang, and Young-Bin Kim. HyCal: A training-free prototype calibration method for cross-discipline few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of CLIP. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33999–34019. PMLR, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 19730–19742. PMLR, 2023.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17612–17625, 2022.

Marco Mistretta, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Andrew D. Bagdanov. Cross the gap: Exposing the intra-modal misalignment in CLIP via modality inversion. In *International Conference on Learning Representations*, 2025.

Junghun Oh, Sungyong Baik, and Kyoung Mu Lee. CLOSER: Towards better representation learning for few-shot class-incremental learning. In *European Conference on Computer Vision*, 2024.

Keon-Hee Park, Kyungwoo Song, and Gyeong-Moon Park. Pre-trained vision and language transformers are few-shot incremental learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23881–23890, 2024.

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. PMLR, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *International Conference on Learning Representations*, 2025.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4077–4087, 2017.

Zeyin Song, Yifan Zhao, Yujun Shi, Peixi Peng, Li Yuan, and Yonghong Tian. Learning with fantasy: Semantic-aware virtual contrastive constraint for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24183–24192, 2023.

Hongbo Sun, Jiahuan Zhou, Xiangteng He, Jinglin Xu, and Yuxin Peng. FineFMPL: Fine-grained feature mining prompt learning for few-shot class incremental learning. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pp. 1299–1307, 2024.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12180–12189, 2020.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pp. 3630–3638, 2016.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Qi-Wei Wang, Da-Wei Zhou, Yi-Kai Zhang, De-Chuan Zhan, and Han-Jia Ye. Few-shot class-incremental learning via training-free prototype calibration. In *Advances in Neural Information Processing Systems*, pp. 18836–18849, 2023.

Wenxuan Wang, Yisi Zhang, Xingjian He, Yichen Yan, Zijia Zhao, Xinlong Wang, and Jing Liu. Beyond literal descriptions: Understanding and locating open-world objects aligned with human intentions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 762–776. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.findings-acl.43. URL https://aclanthology.org/2024.findings-acl.43/.

Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free CLIP-based adaptation. In *International Conference on Learning Representations*, 2024b.

Zongyao Xue, Meina Kan, Shiguang Shan, and Xilin Chen. Feature decomposition-recomposition in large vision-language model for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Shin'ya Yamaguchi, Dewei Feng, Sekitoshi Kanai, Kazuki Adachi, and Daiki Chijiwa. Post-pre-training for modality alignment in vision-language foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4256–4266, 2025.

Yichen Yan, Xingjian He, Sihan Chen, and Jing Liu. Calibration & reconstruction: Deeply integrated language for referring image segmentation. In *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR)*, pp. 451–459. ACM, 2024.

Xi Yang, Wulin Xie, Pai Peng, Jie Wen, and Xiaohuan Lu. Cross-modal mapping: Mitigating the modality gap for few-shot classification. *Pattern Recognition*, 176:113285, 2026. doi: 10.1016/j.patcog.2026.113285.

Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class incremental learning. In *International Conference on Learning Representations*, 2023.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1307–1315, 2018.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12455–12464, 2021.

Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, and De-Chuan Zhan. Forward compatible few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9046–9056, 2022.

A Geometric Derivation and a Source-Matched Counterexample

This appendix gives the derivation and exact synthetic construction underlying Figure 1. The figure illustrates a mathematical counterexample; the measured VLM responses are reported in Section 5.2.

Derivation. Let $a = D_S(q - \mu_S)$ and $b = D_S(r - \mu_S)$ be nonzero. Because D_S is diagonal, $D_S^\top = D_S$ and $M_S = D_S^\top D_S = D_S^2$. Normalizing both vectors gives $a^\top b/(\|a\|_2\|b\|_2)$; substituting M_S yields Eq. (6). The positive query norm is common to every class. Removing this factor from the difference between two scores yields Eq. (7). Thus class ordering depends on both the weighted inner product between query and reference and class-dependent normalization. Reference covariance alone does not encode the distribution of image queries or their correct classes.

Source-matched statistics can reverse a preference. Consider unit references in $\mathbb{R}^2$. A bank of 20 references contains nine copies of e_1, nine copies of $-e_1$, one e_2, and one $-e_2$. Its mean is zero and its unbiased coordinate variances are $(18/19, 2/19)$. Using exactly the paper's calibration with $\rho = 0.1$ gives

$$\widetilde{u}_1 = 17.2/19, \qquad \widetilde{u}_2 = 2.8/19. \tag{12}$$

Take the unit query $q = (2, 1)/\sqrt{5}$ and compare the references e_1 and e_2. Native cosine prefers e_1, since $2/\sqrt{5} > 1/\sqrt{5}$. Let $d_j = 1/\sqrt{\widetilde{u}_j}$. Calibrated and normalized references remain e_1 and e_2, while the calibrated query is proportional to $(2d_1, d_2)$.

$$\frac{d_2}{d_1} = \sqrt{\frac{17.2/19}{2.8/19}} = \sqrt{\frac{43}{7}} > 2. \tag{13}$$

The calibrated query therefore prefers e_2. Its angle to e_1 changes from $\arctan(1/2) \approx 26.6^\circ$ to $\theta' = \arctan\big(d_2/(2d_1)\big) = \arctan\sqrt{43/28} \approx 51.1^\circ$, crossing the 45° equal-cosine boundary shown schematically in Figure 1. The statistics are estimated from the very same reference bank, with the same shrinkage as in the experiments. Because the reversal is strict and the variances are positive, sufficiently small perturbations can make the repeated references distinct while preserving the reversal. The example establishes an ordering reversal under source-matched calibration; the crossed study measures the response across reference roles.

Which geometries preserve native cosine? Zero translation and an isotropic scaling preserve cosine: if $\mu = 0$ and $D^\top D = \lambda I$ for $\lambda > 0$, Eq. (6) reduces to native cosine. The estimated calibration transforms generally translate the origin and rescale directions unequally. Normalizing the transformed features removes their overall length but does not undo these changes in angle. This is why applying one calibration to both endpoints is not itself an invariance argument.

B Complete Incremental Inference Procedure

Table 5: **DS-PoE** inference. All descriptions are fixed.

Input: base training set, incremental support sets, class descriptions, encoders f_I, f_T, and fixed $\rho = 0.1$.
1. Construct base visual prototypes and textual prototypes using Eq. (8).
2. Estimate $(\mu_V, \sqrt{\widetilde{u}_V})$ from base visual prototypes using Eqs. (2)–(3); keep them fixed.
3. Store calibrated base visual references $\bar{v}_c = W_V(v_c)$ and native text references t_c.
4. At each incremental session, average the new support features, normalize the prototypes, and store $\bar{v}_c = W_V(v_c)$. Append visual and fixed text references to their respective banks.
5. For each test image, compute $q = \mathrm{Norm}(f_I(x))$ and $W_V(q)$.
6. Compute $s_c^v = W_V(q)^\top \bar{v}_c$ and $s_c^t = q^\top t_c$ for every known class; predict $\arg\max_c(s_c^v + s_c^t)$.

Product-of-experts identity. Define $p_\ell(c \mid q) = \exp(s_c^\ell)/Z_\ell$ for $\ell \in \{v, t\}$, where $Z_\ell = \sum_j \exp(s_j^\ell)$. Then

$$\frac{p_v(c \mid q)p_t(c \mid q)}{\sum_j p_v(j \mid q)p_t(j \mid q)} = \frac{\exp(s_c^v + s_c^t)}{\sum_j \exp(s_j^v + s_j^t)}. \tag{14}$$

Both Z_v and Z_t cancel algebraically.

Classifier state and complexity. The sufficient inference state contains two reference banks of CD scalars each and two vectors of D calibration statistics. For $C = 1{,}000$ and $D = 512$, this is $1{,}025{,}024$ scalars, or approximately 3.91 MiB in FP32. Diagnostic implementations may additionally retain uncalibrated references. Constructing a new visual reference costs $O(KD)$ after feature extraction; scoring one query costs $O(CD)$.

C Implementation and Evaluation Details

Data schedules. The numbers of evaluated sessions, including the base session, are 9 for CIFAR-100 and miniImageNet and 11 for CUB-200 and IN1K-FSCIL. Base training sets are not reduced to five examples per class. All incremental test sets contain the classes observed up to that session; the class labels and class order are fixed across variants. Query images recurring across sessions are not treated as independent samples in the final-session bootstrap.

Prototype and encoder conventions. The main CLIP implementation averages the unnormalized outputs of the image encoder within each class, then normalizes the class mean. Queries and textual prototypes are normalized separately. The cross-encoder wrappers for SigLIP and EVA-CLIP return normalized per-image features, which are averaged and normalized again to construct their visual references. These conventions are held fixed across every calibration intervention within an encoder.

The CLIP image encoder is evaluated in half precision and its output is cast to FP32 for prototype construction and scoring. Calibration uses the unbiased sample variance across normalized class prototypes before shrinkage. Each encoder uses its own native image preprocessing and text tokenizer. Token length limits are handled by the corresponding tokenizer.

Text inputs. Each class is represented by one fixed string of the form `class name: description`. Descriptions are prepared offline with an LLM following the protocol for class descriptions in FDR (Xue et al., 2025), encoded once, and cached. The same fixed text bank is used by all variants of a given controlled comparison. For the crossed CUB controls, one description record is corrected, and the corrected text bank is shared identically across all paired variants; these runs using the corrected description are reported separately from the standard main result. The corrected record is class ID 84 (zero-based), `085.Horned_Lark`, whose description is "A Horned Lark with a small, compact body and a somewhat flat head." The class-name prefix is retained.

Standard runs and crossed controls. The main method ablation and full matrix of eight variants use the standard saved evaluation artifacts. The crossed study extends these runs with calibration estimated from base textual references and texts available at each session, using paired reruns with the corrected description for CUB. Each intervention response uses its paired native scorer from the same run. The standard matrix is in Table 6; crossed responses and corresponding composed controls are in Tables 1 and 10.

External comparisons. TEEN, FDR, FeCAM, and BiMC in Table 3 are independently reproduced under our shared evaluation protocol. Comparison values are displayed to one decimal place; means across benchmarks are computed before rounding.

D Complete Results for the Scoring Rules

Table 6 provides the eight scoring rules on all 12 encoder–dataset settings. The two isolated visual scorers and two isolated textual scorers identify calibration response. The four composition rules compare native-score addition, shared visual calibration, equal-weight prototype mixing, and source-preserving **DS-PoE**. Every rule within a row uses the same reference banks and test queries.

Across this matrix, **DS-PoE** improves macro A_{avg} by 1.2 points over native-score addition and 1.1 points over shared visual calibration. Its mean advantage over equal-weight prototype mixing is 0.9 points, with higher accuracy in all 12 settings. EVA-CLIP has little change in visual accuracy after calibration, alongside a negative textual response to calibration based on visual references.

Table 6: Complete A_{avg} matrix (%). V_W and T_W use base visual calibration; Native+, Shared+, and Mix denote native-score addition, shared visual calibration, and equal-weight prototype mixing.

Encoder	Dataset	V	V_W	T	T_W	Native+	Shared+	Mix	**DS-PoE**
CLIP	CIFAR-100	67.8	71.9	63.0	60.9	71.1	72.5	72.2	73.1
CLIP	miniImageNet	88.0	91.1	89.5	73.0	90.7	90.5	91.0	92.4
CLIP	CUB-200	73.5	74.6	58.0	46.7	73.7	72.8	74.3	76.0
CLIP	IN1K-FSCIL	66.3	70.4	66.1	44.3	71.3	70.5	71.4	72.7
SigLIP	CIFAR-100	72.5	74.9	67.2	66.8	75.8	76.3	75.8	76.1
SigLIP	miniImageNet	93.5	94.7	91.3	75.9	94.9	95.2	94.8	95.4
SigLIP	CUB-200	76.7	76.6	58.2	43.4	74.7	75.2	75.7	77.3
SigLIP	IN1K-FSCIL	76.9	78.6	71.5	48.7	78.9	78.8	79.3	79.8
EVA-CLIP	CIFAR-100	88.2	88.6	84.6	83.2	88.7	89.3	88.8	89.3
EVA-CLIP	miniImageNet	93.0	93.1	92.2	90.1	94.0	94.3	94.1	94.5
EVA-CLIP	CUB-200	81.1	80.9	63.7	57.7	80.3	80.1	80.2	81.3
EVA-CLIP	IN1K-FSCIL	76.8	76.9	72.8	70.1	78.7	78.5	78.5	79.0

Table 7: CLIP expert complementarity. Percentages are averaged equally across sessions. Visual and text denote the calibrated visual and native textual experts. Columns 2–5 partition the queries.

Dataset	Both correct	Visual only	Text only	Both wrong	**DS-PoE**
CIFAR-100	54.7	17.3	8.4	19.7	73.1
miniImageNet	84.8	6.3	4.8	4.1	92.4
CUB-200	50.2	24.4	7.9	17.5	76.0
IN1K-FSCIL	56.4	14.0	9.7	19.9	72.7

Table 7 partitions predictions by which expert is correct. Both experts make unique correct decisions on every CLIP dataset. Adding the proportions for the *Both correct*, *Visual only*, and *Text only* columns gives the oracle union; **DS-PoE** makes its decisions from the two score vectors.

E Paired Bootstrap over Classes

For each class c at the final session, let $a_c^{R,S}$ denote its accuracy under scorer (R, S). We form the paired class-level contrast

$$g_c = (a_c^{V,V} - a_c^{V,I}) - (a_c^{T,V} - a_c^{T,I}), \qquad \Gamma_{\text{final}} = \frac{1}{C}\sum_{c=1}^{C} g_c. \tag{15}$$

Each bootstrap replicate samples C classes with replacement and averages their g_c values. All four sets of predictions for a class remain paired. The 2.5th and 97.5th percentiles of 10,000 draws define the interval. Table 8 reports this class-balanced point estimate and its interval for the final session. In particular, these intervals are not intervals for the mean Γ across sessions in Eq. (5). This distinction also accounts for unequal numbers of query images per CUB class.

F Definition of the Ordering Response

Let $C_t = |\mathcal{C}_{\le t}|$ and let N_t be the number of test queries in session t. For query i with label y_i, let

$$B_t(s_i, y_i) = \frac{1}{C_t - 1}\sum_{c \neq y_i} \mathbf{1}[s_{i,y_i} > s_{i,c}]. \tag{16}$$

For an isolated scorer before and after a calibration, the reported response is

$$\Omega = \frac{100}{T+1}\sum_{t=0}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}\left[B_t(s_i^{\text{calibrated}}, y_i) - B_t(s_i^{\text{native}}, y_i)\right]. \tag{17}$$

Table 8: Paired bootstrap over classes for the class-balanced interaction Γ_{final} at the final session (percentage points; 10,000 resamples). Point estimates and intervals use the same statistic.

Encoder	Dataset	Γ_{final}	95% interval
CLIP	CIFAR-100	+7.0	[+2.4, +11.8]
CLIP	miniImageNet	+20.5	[+16.0, +25.2]
CLIP	CUB-200	+12.7	[+8.4, +17.1]
CLIP	IN1K-FSCIL	+26.3	[+24.5, +28.2]
SigLIP	CIFAR-100	+3.4	[-0.7, +7.5]
SigLIP	miniImageNet	+18.9	[+14.0, +23.9]
SigLIP	CUB-200	+16.6	[+12.5, +20.8]
SigLIP	IN1K-FSCIL	+24.9	[+22.9, +26.9]
EVA-CLIP	CIFAR-100	+2.9	[+0.2, +5.7]
EVA-CLIP	miniImageNet	+2.9	[+0.9, +5.1]
EVA-CLIP	CUB-200	+5.9	[+2.5, +9.4]
EVA-CLIP	IN1K-FSCIL	+3.0	[+2.0, +4.0]

Table 9: Absolute A_{avg} (%) for the paired native and calibrated cells underlying Table 1.

Encoder	Dataset	V, I	T, I	V, V	T, V	V, T_b	T, T_b	V, T_t	T, T_t
CLIP	CIFAR-100	67.8	63.1	71.9	60.9	63.7	56.5	64.0	56.8
CLIP	miniImageNet	88.0	89.5	91.1	73.0	82.2	83.0	82.4	84.4
CLIP	CUB-200	73.5	57.9	74.6	46.8	66.1	6.1	66.4	6.8
CLIP	IN1K-FSCIL	66.2	66.2	70.4	44.3	60.4	52.9	60.4	52.7
SigLIP	CIFAR-100	72.5	67.2	74.9	66.8	71.7	15.8	71.7	13.5
SigLIP	miniImageNet	93.5	91.3	94.7	75.9	93.0	34.6	93.1	39.4
SigLIP	CUB-200	76.7	58.2	76.6	43.3	74.7	6.6	74.8	8.6
SigLIP	IN1K-FSCIL	76.9	71.5	78.6	48.7	76.7	4.2	76.7	4.2
EVA-CLIP	CIFAR-100	88.2	84.6	88.6	83.2	88.2	65.5	88.2	66.3
EVA-CLIP	miniImageNet	93.0	92.2	93.1	90.1	93.0	79.5	93.0	78.7
EVA-CLIP	CUB-200	81.1	63.6	80.9	57.7	81.2	47.3	81.1	48.6
EVA-CLIP	IN1K-FSCIL	76.8	72.8	76.9	70.1	76.9	49.4	76.9	49.4

Equivalently, within each session we count helpful transitions from a loss or tie to a win and harmful transitions in the other direction, divide their difference by $N_t(C_t - 1)$, and multiply by 100 to express the response in percentage points. Sessions receive equal weight regardless of their query counts or numbers of competing classes.

The analysis uses stored CLIP scores cast to FP32. Exact ties contribute zero under the strict comparison above. Accuracy and bootstrap results use saved predictions.

G Composition Controls with Matched Statistics

Table 9 reports native and calibrated accuracies for all isolated scorers in the crossed study.

The interventions using textual statistics suggest a natural control for the combined classifier: keep the calibrated visual expert and replace native text with a scorer using its own statistics. We evaluate

$$s_c^{\text{matched},b} = s_c^{V,V} + s_c^{T,T_b}, \qquad s_c^{\text{matched},t} = s_c^{V,V} + s_c^{T,T_t}. \tag{18}$$

DS-PoE uses $s_c^{V,V} + s_c^{T,I}$. All three share the visual expert and fixed addition. Preserving native textual scoring yields the highest macro accuracy and outperforms both controls with matched textual statistics in 10/12 settings. On CLIP, **DS-PoE** reaches 78.56%, versus 77.83% and 77.79% for the controls using base and current textual statistics, respectively.

Table 10: Composition controls for textual calibration: native text in **DS-PoE** versus text calibrated with base or current textual statistics. The visual expert uses W_V throughout. Values are A_{avg} (%).

Encoder	Dataset	**DS-PoE**	Matched base	Matched current
CLIP	CIFAR-100	73.1	73.2	73.2
CLIP	miniImageNet	92.4	92.3	92.3
CLIP	CUB-200	76.0	73.9	73.7
CLIP	IN1K-FSCIL	72.7	71.9	71.9
SigLIP	CIFAR-100	76.1	74.4	74.1
SigLIP	miniImageNet	95.4	93.4	93.5
SigLIP	CUB-200	77.3	65.7	66.8
SigLIP	IN1K-FSCIL	79.8	74.5	74.5
EVA-CLIP	CIFAR-100	89.3	89.3	89.3
EVA-CLIP	miniImageNet	94.5	94.3	94.3
EVA-CLIP	CUB-200	81.3	79.6	79.5
EVA-CLIP	IN1K-FSCIL	79.0	77.4	77.3

Table 11: Robustness to resampling support images with three seeds: A_{avg} mean $\pm$ sample standard deviation (%). Base data, test sets, and class order are fixed; incremental supports are resampled.

Scoring rule	CIFAR	miniIN	CUB	IN1K
Visual, native	67.47 ± 0.34	87.67 ± 0.34	73.24 ± 0.48	66.26 ± 0.06
Visual, calibrated	71.80 ± 0.41	90.70 ± 0.25	74.19 ± 0.24	70.46 ± 0.06
Text, native	63.05 ± 0.00	89.55 ± 0.00	58.04 ± 0.00	66.15 ± 0.00
Text, visually calibrated	60.85 ± 0.00	72.98 ± 0.00	46.76 ± 0.00	44.33 ± 0.00
Native-score addition	71.45 ± 0.16	90.36 ± 0.16	73.60 ± 0.24	70.99 ± 0.02
Shared visual calibration	72.34 ± 0.27	90.33 ± 0.16	72.99 ± 0.30	70.34 ± 0.02
Equal-weight prototype mixing	71.94 ± 0.05	90.90 ± 0.17	74.24 ± 0.06	71.32 ± 0.07
DS-PoE	73.12 ± 0.37	92.08 ± 0.13	75.77 ± 0.23	72.76 ± 0.03

H Support Resampling and Shot Sensitivity

For three seeds, incremental supports are sampled from each class's full training pool while holding the base data, class order, and evaluation images fixed. The same sampled supports are shared across variants. We report the sample standard deviation across the three resulting A_{avg} values. Native textual scores and calibration of textual scoring using visual statistics remain fixed when support images are resampled because their text bank and base visual statistics are unchanged, giving zero standard deviations.

The calibrated visual expert improves over the native visual expert on all four datasets under these resampled supports. **DS-PoE** also exceeds native-score addition in all four dataset means. Table 11 retains the full variant comparison, distinguishing standalone accuracy from composition outcomes.

The shot sweep varies only incremental support size, using $K \in \{1, 2, 5, 10\}$ on CUB-200, miniImageNet, and IN1K-FSCIL. Table 12 shows how visual estimation responds to support size while the semantic anchor remains fixed. On miniImageNet, **DS-PoE** trails native text at one shot but surpasses it as more visual support becomes available. On IN1K-FSCIL, **DS-PoE** improves from 70.1% at one shot to 73.5% at ten shots and remains above either individual expert at each support size.

I Calibration Compatibility across Shrinkage Strengths

We vary only the shrinkage coefficient for diagonal variances $\rho \in \{0, 0.05, 0.1, 0.2, 0.5, 1\}$ on all four benchmarks. For each value, the same three seeds for resampling support images and paired native scorers determine $\Delta_V = \Delta_{V \leftarrow V}$ and $\Delta_T = \Delta_{T \leftarrow V}$ as defined in Eq. (5); $\Gamma = \Delta_V - \Delta_T$. The displayed **DS-PoE** standard deviation is the sample standard deviation across those seeds. Native

Table 12: Effect of the number of incremental shots with three support seeds: A_{avg} mean $\pm$ sample standard deviation (%). Five-shot rows use the reference runs in Table 11. Base session data remain fixed. Native+ adds native visual and textual scores.

Dataset	Shots	V	V_W	T	Native+	**DS-PoE**
CUB-200	1	61.29±0.11	64.94±0.16	58.04±0.00	67.09±0.17	67.43±0.21
CUB-200	2	66.99±0.60	69.43±0.24	58.04±0.00	71.23±0.13	71.62±0.26
CUB-200	5	73.24±0.48	74.19±0.24	58.04±0.00	73.60±0.24	75.77±0.23
CUB-200	10	75.79±0.15	76.63±0.26	58.04±0.00	75.25±0.09	77.81±0.17
miniImageNet	1	76.56±1.08	83.87±0.67	89.55±0.00	83.91±0.74	87.62±0.64
miniImageNet	2	82.29±1.09	88.21±0.35	89.55±0.00	89.55±0.29	90.53±0.13
miniImageNet	5	87.67±0.34	90.70±0.25	89.55±0.00	90.36±0.16	92.08±0.13
miniImageNet	10	88.54±0.14	91.07±0.07	89.55±0.00	92.29±0.09	92.39±0.07
IN1K-FSCIL	1	63.29±0.11	67.71±0.30	66.15±0.00	68.57±0.24	70.09±0.29
IN1K-FSCIL	2	64.50±0.16	68.98±0.10	66.15±0.00	70.13±0.12	71.34±0.12
IN1K-FSCIL	5	66.26±0.06	70.46±0.06	66.15±0.00	70.99±0.02	72.76±0.03
IN1K-FSCIL	10	67.29±0.03	71.42±0.06	66.15±0.00	72.47±0.04	73.51±0.08

Table 13: CLIP calibration strength across three support seeds. **DS-PoE** is mean $A_{\text{avg}}\pm$ sample standard deviation (%); responses and Γ are mean percentage-point differences relative to paired native scorers. The paper uses $\rho = 0.1$.

Dataset	ρ	**DS-PoE**	Δ_V	Δ_T	Γ
CIFAR-100	0	73.17±0.34	+4.23	-5.28	+9.51
CIFAR-100	0.05	73.16±0.37	+4.32	-3.48	+7.81
CIFAR-100	**0.1**	73.12±0.37	+4.32	-2.20	+6.52
CIFAR-100	0.2	73.04±0.35	+4.32	-0.61	+4.94
CIFAR-100	0.5	72.78±0.28	+4.00	+1.79	+2.21
CIFAR-100	1	70.26±0.42	+0.64	+2.86	-2.23
miniImageNet	0	92.08±0.16	+3.01	-34.45	+37.46
miniImageNet	0.05	92.11±0.13	+3.01	-23.17	+26.18
miniImageNet	**0.1**	92.08±0.13	+3.02	-16.56	+19.59
miniImageNet	0.2	92.11±0.15	+2.99	-10.49	+13.48
miniImageNet	0.5	92.00±0.17	+2.63	-5.20	+7.82
miniImageNet	1	90.48±0.21	+0.46	-2.14	+2.60
CUB-200	0	75.80±0.23	+0.77	-30.51	+31.28
CUB-200	0.05	75.76±0.18	+0.92	-15.46	+16.38
CUB-200	**0.1**	75.77±0.23	+0.94	-11.27	+12.21
CUB-200	0.2	75.77±0.19	+0.97	-6.77	+7.74
CUB-200	0.5	75.72±0.15	+0.77	-3.00	+3.77
CUB-200	1	75.08±0.22	-0.16	-1.85	+1.69
IN1K-FSCIL	0	72.77±0.03	+4.24	-35.32	+39.56
IN1K-FSCIL	0.05	72.74±0.04	+4.21	-27.04	+31.26
IN1K-FSCIL	**0.1**	72.76±0.03	+4.21	-21.83	+26.05
IN1K-FSCIL	0.2	72.68±0.04	+4.13	-15.40	+19.53
IN1K-FSCIL	0.5	72.35±0.03	+3.54	-8.07	+11.62
IN1K-FSCIL	1	69.83±0.05	-0.30	-4.95	+4.65

textual scoring and its calibration response do not depend on incremental support samples, so the textual response is constant across seeds.

From $\rho = 0$ through 0.2, **DS-PoE** varies by at most 0.13 percentage points within each dataset, while $\Delta_V > 0$, $\Delta_T < 0$, and $\Gamma > 0$ on all four. Thus the chosen $\rho = 0.1$ lies in a broad stable range rather than at a narrow accuracy peak. Stronger shrinkage can change the response signs on CIFAR-100; at $\rho = 1$, the per-coordinate scale becomes uniform and cancels under normalization, leaving centering-only calibration.

J SENSITIVITY TO PROTOTYPE MIXING

Table 14: Sensitivity to the prototype mixing coefficient. Oracle α maximizes evaluation accuracy over the displayed grid. Mean $\pm$ sample standard deviation over three support samples.

Dataset	**DS-PoE**	Fixed $\alpha = 0.5$	Oracle α	Oracle result
CIFAR-100	73.12 ± 0.37	71.94 ± 0.05	0.50	71.94 ± 0.05
miniImageNet	92.08 ± 0.13	90.90 ± 0.17	0.25	92.76 ± 0.09
CUB-200	75.77 ± 0.23	74.24 ± 0.06	0.75	76.07 ± 0.26
IN1K-FSCIL	72.76 ± 0.03	71.32 ± 0.07	0.25	71.92 ± 0.00

The mixing control constructs

$$m_c(\alpha) = \mathrm{Norm}(\alpha v_c + (1-\alpha)t_c), \qquad s_c^{\mathrm{mix}}(q) = q^\top m_c(\alpha). \tag{19}$$

We inspect $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ on the three support seeds. The best coefficient in Table 14 is selected retrospectively from evaluation accuracy and serves as an oracle diagnostic. The preferred coefficient varies across the evaluated benchmarks. These oracle results characterize the sensitivity of prototype interpolation, rather than a deployable tuned baseline.

K LATE SCORE COMPOSITION SENSITIVITY

Table 15: Sensitivity to score weighting on CLIP ViT-B/16. Here λ is the weight on the visual score in $s_c^\lambda = \lambda s_c^v + (1-\lambda)s_c^t$; $\lambda = 0.5$ is equivalent to the fixed **DS-PoE** rule under $\arg\max$. Values are mean session accuracy (%). The text-only endpoint uses Table 11; the other rows retain the canonical evaluation used in Table 3.

λ (visual weight)	CIFAR-100	miniImageNet	CUB-200	IN1K-FSCIL	Mean
0.00 (text only)	63.05	89.55	58.04	66.15	69.20
0.25	**73.58**	92.16	75.93	**72.87**	**78.64**
0.50 (**DS-PoE**)	73.09	**92.44**	**76.01**	72.73	78.57
0.75	72.50	91.85	75.24	71.53	77.78
1.00 (visual only)	71.92	91.11	74.64	70.43	77.02

DS-PoE combines the calibrated visual and native textual experts through fixed score addition,

$$s_c(q) = s_c^v(q) + s_c^t(q). \tag{20}$$

To examine the relative contribution of the two score vectors without changing either reference bank, we evaluate the controlled composition

$$s_c^\lambda(q) = \lambda s_c^v(q) + (1-\lambda)s_c^t(q), \tag{21}$$

for $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$ on CLIP ViT-B/16 across all four benchmarks. The endpoints give the isolated textual and visual experts, while $\lambda = 0.5$ has the same class ordering as **DS-PoE**.

Among the tested ratios, $\lambda = 0.25$ yields the numerically highest mean across the four benchmarks (78.64% versus 78.57% for $\lambda = 0.5$), whereas miniImageNet and CUB-200 favor equal weighting. The 0.07-point difference and the different preferred weights across datasets support equal addition as a strong fixed rule without dataset-specific tuning.