MASS-PRESERVING VISUAL–TEXT FUSION FOR TRAINING-FREE FEW-SHOT CLASS-INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Training-Free Few-Shot Class-Incremental Learning (FSCIL) must use scarce support images without confusing class-specific evidence with probability reallocation between base and novel classes. This distinction matters because a globally normalized interface can change both the ranking within the novel group and the competition between novel and base classes at the same time. Direct support-image fusion therefore makes the source of an endpoint change difficult to identify. We introduce a Mass-Preserving Visual–Text Fusion (MP-VTF) correction at the Bi-level Modality Calibration (BiMC) interface to separate these effects. MP-VTF uses labeled support images during frozen inference, without updating model parameters or accessing query labels. It applies a query-wise log-sum-exp translation that retains the fused within-novel ordering while restoring the text branch's original novel-group probability mass. Across CIFAR100, CUB-200-2011, and miniImageNet, Naive fusion increased the mean novel-group probability mass in all 78 session–seed records, whereas MP-VTF restored it to numerical precision. MP-VTF produced small positive novel-class changes, nearly unchanged overall accuracy, and a small base-class cost. The miniImageNet endpoint direction is reported descriptively because its visual weight was selected with label-dependent metrics. These results establish mass preservation as a testable interface-control principle for separating within-group evidence updates from group-level probability allocation.

1 INTRODUCTION

Few-Shot Class-Incremental Learning (FSCIL) requires a model to recognize classes learned in a data-rich base session while incorporating novel classes that arrive later with only a few labeled examples (Tao et al., 2020). Each incremental decision must preserve recognition of the accumulated base set while exploiting scarce evidence for new classes. This stability–plasticity requirement is particularly restrictive in training-free settings: a pre-trained vision–language model can provide semantic priors, but its parameters remain frozen as the class set grows (Wang et al., 2023; Chen et al., 2025; Lee et al., 2026). At the BiMC inference interface studied here, the specific question is how labeled novel support images can alter distinctions among novel classes without also changing the text branch's aggregate allocation between the base and novel groups.

Training-free vision–language classification combines complementary evidence sources. Text prototypes and category descriptions provide semantic coverage beyond the few observed images (Radford et al., 2021; Pratt et al., 2023), whereas support-image caches, visual prototypes, and cross-modal fusion provide task-specific instance evidence (Zhang et al., 2022; Zhu et al., 2023; Li et al., 2025). Bi-level Modality Calibration (BiMC) combines calibrated modality-specific predictions in a frozen inference pipeline and uses a text-based branch for accumulated novel classes (Chen et al., 2025). We call direct support-image fusion at this interface without a group-level constraint Naive fusion. Under BiMC's all-class softmax, direct fusion can rerank novel classes while also shifting the total probability assigned to the novel group.

Mass-Preserving Visual–Text Fusion (MP-VTF) separates these effects at BiMC's novel interface. MP-VTF first interpolates Text-CMNN novel logits with Visual 1NN scores from labeled novel

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.

support images. It then applies a query-wise uniform log-sum-exp translation before the all-class softmax. In exact arithmetic, this translation preserves the pairwise differences among the fused novel logits while restoring the original Text-CMNN novel-group probability mass. MP-VTF performs no parameter update and uses no query label during inference.

We compare matched BiMC, Naive fusion, and MP-VTF paths over three seeds on CIFAR100, CUB-200-2011 (CUB200), and miniImageNet. The experiments measure the mass shift, correction residual, and downstream accuracy. Figure 1 locates the correction within BiMC.

Our contributions are threefold:

1. We formalize the coupling between within-novel logit differences and Text-CMNN novel-group probability mass at the globally normalized BiMC interface.
2. We introduce a query-wise log-sum-exp translation that preserves the fused within-novel ordering and restores the Text-CMNN novel-group probability mass.
3. We provide controlled evidence across three FSCIL benchmarks, including numerical constraint satisfaction, matched endpoint comparisons, and session-wise analyses.

2 Related Work

2.1 FSCIL under Limited Data

In the FSCIL setting introduced by Tao et al. [2020], most methods adapt a representation or classifier during training. Examples include prototype refinement, relation distillation, meta-learning, and multi-phase task sampling (Zhu et al., 2021; Dong et al., 2021; Chi et al., 2022; Zhou et al., 2023). Continually Evolved Classifiers (CEC) learns a graph module on pseudo-incremental tasks to adapt class weights across sessions (Zhang et al., 2021). Forward Compatible Training (FACT) learns virtual prototypes that reserve space for future classes (Zhou et al., 2022). Other approaches use computational constraints, bilateral distillation, multi-grained update schedules, or synthesized-feature subspaces (Hersche et al., 2022; Zhao et al., 2023; 2024; Cheraghian et al., 2021). Semantic-Aware Virtual Contrastive (SAVC) learning constructs virtual classes and optimizes contrastive constraints to separate base from prospective novel classes (Song et al., 2023). Each addresses FSCIL through a learned representation, module, or objective; our question concerns a frozen inference interface.

2.2 Training-Free FSCIL with Vision–Language Models

Pre-trained vision–language models introduce linguistic priors into FSCIL, although many such methods remain training-dependent. Continual Parameter-Efficient CLIP (CPE-CLIP) learns visual and language prompts across sessions (D'Alessandro et al., 2023), and PriViLege tunes prompts and selected transformer layers in the base session while transferring semantic knowledge by distillation (Park et al., 2024). Training-free approaches instead modify stored representations or inference rules without updating model parameters. TEEN recalibrates a novel visual prototype by mixing it with semantically related base prototypes after the feature extractor has been trained (Wang et al., 2023). BiMC goes further toward a fully frozen vision–language setting: it treats CLIP as a black-box model, calibrates text and visual evidence, and combines modality-specific predictions through Masked Ensemble inference (Chen et al., 2025). Recent work has also explored diffusion-generated prototypes and hybrid-distance prototype calibration under cross-discipline shifts (Kang et al., 2025; Lee et al., 2026), but these interventions occur during prototype construction or prototype-based scoring rather than at BiMC's globally normalized text-based novel interface.

2.3 Support-Image Fusion at Globally Normalized Interfaces

CLIP matches an image to text prototypes built from class prompts (Radford et al., 2021). CuPL enriches fixed text classifiers with generated category descriptions (Pratt et al., 2023), whereas Saha et al. use descriptions with image collections for learned bag-level image–text supervision (Saha et al., 2024). BiMC draws on generated descriptions in a frozen text branch; MP-VTF leaves that semantic evidence unchanged and constrains only how support-image evidence modifies its novel-group probability mass. Training-free few-shot adaptation instead uses labeled support images through caches or visual prototypes: Tip-Adapter adds retrieved label evidence to the text logits

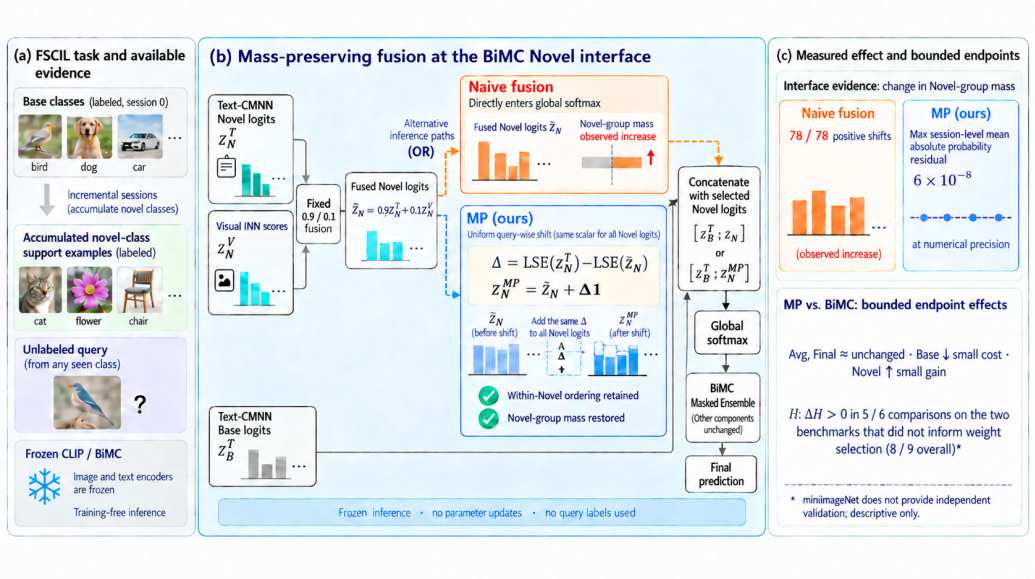

Figure 1: MP-VTF at the BiMC novel interface. (a) The base/novel/query split and frozen inference setting. (b) The Naive and MP-VTF paths, where Δ is the uniform log-sum-exp shift that restores the Text-CMNN novel-group probability mass while preserving the fused within-novel ordering. (c) Measured interface behavior and endpoint directions. Base logits, the fused-prototype branch, the ensemble weight, and the prediction rule are unchanged.

(Zhang et al., 2022), APE reweights the cache from query–support–text relations (Zhu et al., 2023), TIMO combines text- and image-guided classifiers (Li et al., 2025), and a recent preprint ensembles a text-aligned prototype classifier with image-space linear discriminant analysis scores (Goswami et al., 2026). These methods study conventional few-shot classification rather than an accumulated base–novel interface. Class-incremental learning has addressed old–new score bias with a learned bias layer (Wu et al., 2019) or a separated softmax objective (Ahn et al., 2021). Temperature scaling recalibrates probabilities post hoc without retraining (Guo et al., 2017), but it applies a global scalar and does not isolate a group-internal update from group-level allocation. MP-VTF targets that narrower interface question and leaves the remaining BiMC components intact.

3 Method

Direct support-image fusion at BiMC's novel interface entangles within-novel logit differences with changes in the Text-CMNN novel-group probability mass. MP-VTF separates these quantities through a query-wise log-sum-exp correction. Here, novel-group probability mass means the sum of the novel-class probabilities after the Text-CMNN all-class softmax and before the Masked Ensemble. Figure 1 shows the controlled interface and its location within the unchanged BiMC pipeline.

3.1 Problem Setup and BiMC Interface

In FSCIL, a base session is followed by incremental sessions that introduce previously unseen classes from only a few labeled examples. At session t the seen-class set is $\mathcal{C}^{(t)} = \mathcal{C}_{\mathrm{B}} \cup \mathcal{C}_{\mathrm{N}}^{(t)}$; given the cumulative labeled novel support set $\mathcal{S}_{\mathrm{N}}^{(t)}$ and an unlabeled query image x, the objective is to predict a class in $\mathcal{C}^{(t)}$. CLIP ViT-B/16 remains frozen as the class set grows, and q denotes its image feature for x. The expressions below describe one query; the implementation evaluates queries in batches.

Bi-level Modality Calibration (BiMC) (Chen et al., 2025) uses a cross-modal category nearest-neighbor metric (CMNN), called Text-CMNN here to distinguish it from Visual 1NN (Section 3.2). Its Masked Ensemble combines fused-prototype and Mahalanobis probabilities for base classes, and fused-prototype and Text-CMNN probabilities for novel classes. Let z^{T} denote the Text-CMNN

logits over all seen classes, with base slice $z_{\mathrm{B}}^{\mathrm{T}}$ and novel slice $z_{\mathrm{N}}^{\mathrm{T}}$, and $p^{\mathrm{T}} = \mathrm{softmax}(z^{\mathrm{T}})$ with novel slice $p_{\mathrm{N}}^{\mathrm{T}}$. With $p_{\mathrm{N}}^{\mathrm{proto}}$ the fused-prototype novel probabilities and α BiMC's ensemble weight, the BiMC novel score vector is

$$s_{\mathrm{N}}^{\mathrm{BiMC}} = \alpha p_{\mathrm{N}}^{\mathrm{proto}} + (1-\alpha) p_{\mathrm{N}}^{\mathrm{T}}.$$

MP-VTF modifies only the Text-CMNN component of this novel branch, retaining the Text-CMNN base logits, fused-prototype probabilities, ensemble weight α, and final base branch of the same BiMC inference. At the base session ($t = 0$), $\mathcal{C}_{\mathrm{N}}^{(0)} = \varnothing$, so there are no novel logits to modify and MP-VTF is identical to BiMC. During incremental sessions it uses the labeled novel support images to score the accumulated novel classes and replaces $p_{\mathrm{N}}^{\mathrm{T}}$ with a corrected novel-class probability vector.

3.2 Naive Visual–Text Fusion

We first construct a visual score from the labeled support images of the novel classes accumulated up to session t. For each novel class $c \in \mathcal{C}_{\mathrm{N}}^{(t)}$, let $\mathcal{S}_c$ be the set of support-image features assigned to that class. After ℓ_2 normalization, the Visual 1NN score is

$$z_{\mathrm{N},c}^{\mathrm{V}} = \max_{u \in \mathcal{S}_c} \langle \bar{q}, \bar{u} \rangle.$$

The support labels are used only to organize the available few-shot examples by class; the query label is not used. We combine the resulting novel-class score vector $z_{\mathrm{N}}^{\mathrm{V}}$ with the Text-CMNN novel logits $z_{\mathrm{N}}^{\mathrm{T}}$ using a fixed visual weight $w = 0.1$:

$$\widetilde{z}_{\mathrm{N}} = (1-w) z_{\mathrm{N}}^{\mathrm{T}} + w z_{\mathrm{N}}^{\mathrm{V}}.$$

Both $z_{\mathrm{N}}^{\mathrm{T}}$ and $z_{\mathrm{N}}^{\mathrm{V}}$ are pre-softmax cosine-similarity scores that the fusion interface treats as logits. This direct interpolation couples two effects. It can change the relative ordering of the novel classes, which is the intended role of the visual evidence, but it generally also changes their aggregate log-sum-exp (LSE):

$$\mathrm{LSE}(\widetilde{z}_{\mathrm{N}}) \neq \mathrm{LSE}(z_{\mathrm{N}}^{\mathrm{T}}), \qquad \mathrm{LSE}(z) = \log \sum_j \exp(z_j).$$

Because the base logits remain unchanged and the subsequent softmax is evaluated over all seen classes, this log-sum-exp drift changes the novel-group probability mass. Naive fusion therefore couples fused within-novel reranking to a shift in that mass. MP-VTF keeps the former while analytically correcting the latter.

3.3 Mass-Preserving Fusion

MP-VTF corrects the Naive novel logits with a single additive shift; the equations below write VTF for MP-VTF. For each query, we compute

$$\Delta = \mathrm{LSE}(z_{\mathrm{N}}^{\mathrm{T}}) - \mathrm{LSE}(\widetilde{z}_{\mathrm{N}})$$

and add this scalar uniformly to every accumulated novel-class logit:

$$z_{\mathrm{N}}^{\mathrm{VTF}} = \widetilde{z}_{\mathrm{N}} + \Delta \mathbf{1},$$

where $\mathbf{1}$ is an all-ones vector with the same dimension as $\widetilde{z}_{\mathrm{N}}$. The shift restores the original Text-CMNN novel log-sum-exp in exact arithmetic:

$$\begin{aligned}\mathrm{LSE}(z_{\mathrm{N}}^{\mathrm{VTF}}) &= \mathrm{LSE}(\widetilde{z}_{\mathrm{N}} + \Delta\mathbf{1}) \\ &= \mathrm{LSE}(\widetilde{z}_{\mathrm{N}}) + \Delta \\ &= \mathrm{LSE}(z_{\mathrm{N}}^{\mathrm{T}}).\end{aligned}$$

We then retain the original Text-CMNN base logits $z_{\mathrm{B}}^{\mathrm{T}}$ and apply a global softmax over all seen classes, where $[a; b]$ denotes vector concatenation:

$$p^{\mathrm{VTF}} = \mathrm{softmax}\big([z_{\mathrm{B}}^{\mathrm{T}}; z_{\mathrm{N}}^{\mathrm{VTF}}]\big)\,, \qquad p^{\mathrm{T}} = \mathrm{softmax}\big([z_{\mathrm{B}}^{\mathrm{T}}; z_{\mathrm{N}}^{\mathrm{T}}]\big)\,.$$

Naive fusion corresponds instead to the all-class probability vector $p^{\mathrm{Naive}} = \mathrm{softmax}\left([z_{\mathrm{B}}^{\mathrm{T}}; \widetilde{z}_{\mathrm{N}}]\right)$. We define the per-query novel-group probability shift induced by the unconstrained interface as the signed difference

$$\delta^{\mathrm{Naive}} = \sum_{c \in \mathcal{C}_{\mathrm{N}}^{(t)}} p_c^{\mathrm{Naive}} - \sum_{c \in \mathcal{C}_{\mathrm{N}}^{(t)}} p_c^{\mathrm{T}},$$

A positive value means that Naive fusion assigns more probability mass to the novel group than Text-CMNN; Section 4.2 reports the measured sign. MP-VTF leaves the base logits unchanged and matches Text-CMNN's summed exponentiated novel logits. Its corresponding shift $\delta^{\mathrm{VTF}} = \sum_{c \in \mathcal{C}_{\mathrm{N}}^{(t)}} \left(p_c^{\mathrm{VTF}} - p_c^{\mathrm{T}}\right)$ is therefore zero in exact arithmetic:

$$\sum_{c \in \mathcal{C}_{\mathrm{N}}^{(t)}} p_c^{\mathrm{VTF}} = \sum_{c \in \mathcal{C}_{\mathrm{N}}^{(t)}} p_c^{\mathrm{T}}.$$

Finally, MP-VTF changes only the Text-CMNN term in BiMC's novel Masked Ensemble. Let $p_{\mathrm{N}}^{\mathrm{proto}}$ denote the unchanged fused-prototype probabilities for the novel classes, and let $p_{\mathrm{N}}^{\mathrm{VTF}}$ be the novel-class slice of the global softmax p^{VTF}. Using the original BiMC ensemble weight α, the reconstructed novel scores are

$$s_{\mathrm{N}}^{\mathrm{VTF}} = \alpha p_{\mathrm{N}}^{\mathrm{proto}} + (1 - \alpha) p_{\mathrm{N}}^{\mathrm{VTF}}.$$

Thus, the complete MP-VTF forward path is Visual 1NN scoring, fixed novel-logit interpolation, a query-wise LSE shift, global softmax over all seen classes, and reconstruction of the novel Masked Ensemble. BiMC's unchanged base scores and the reconstructed novel scores are then compared jointly, and the final prediction is the class with the maximum ensemble score over all seen classes.

3.4 Interface Invariances of MP-VTF

MP-VTF provides two exact invariances at the fusion interface. Because the same scalar Δ is added to every Naive novel logit, the pairwise differences satisfy

$$z_{\mathrm{N},i}^{\mathrm{VTF}} - z_{\mathrm{N},j}^{\mathrm{VTF}} = \widetilde{z}_{\mathrm{N},i} - \widetilde{z}_{\mathrm{N},j}.$$

MP-VTF therefore preserves the fused within-novel ordering produced by Naive fusion while restoring the Text-CMNN novel-group probability mass. Mass preservation, together with the unchanged $p_{\mathrm{N}}^{\mathrm{proto}}$ and α, further implies that the sum of the novel Masked Ensemble scores is retained relative to BiMC:

$$\mathbf{1}^{\top} s_{\mathrm{N}}^{\mathrm{VTF}} = \mathbf{1}^{\top} s_{\mathrm{N}}^{\mathrm{BiMC}}.$$

We evaluate the resulting all-class predictions empirically in Section 4.

Table 1: Evaluation protocols and frozen inference settings

Dataset	base classes	Incremental sessions	Classes/session	Shots/class	Total sessions	Ensemble α
CIFAR100	60	8	5	5	9	0.60
CUB200	100	10	5	5	11	0.80
miniImageNet	60	8	5	5	9	0.60

All datasets use the BiMC FSCIL protocol, three random seeds, a frozen CLIP ViT-B/16 encoder, a fixed visual weight of $w = 0.1$, and the official pre-generated category descriptions. Total sessions include the base session.

3.5 Training-Free Inference and Label Use

MP-VTF adds no optimization stage to the frozen BiMC pipeline. The CLIP encoders and all BiMC components remain fixed. MP-VTF uses normalized feature multiplication, class-wise maxima over support similarities, fixed logit interpolation, an LSE shift, and a global softmax. The visual weight $w = 0.1$ was selected in a preliminary miniImageNet seed-1 study using Avg, Final, and H, then fixed for all nine reported runs. The same value was applied to all three datasets. MP-VTF inherited BiMC's dataset-specific ensemble weights without reselection: $\alpha = 0.60$ for CIFAR100 and miniImageNet and $\alpha = 0.80$ for CUB200.

After w has been fixed, MP-VTF uses labeled support images and unlabeled query features. Query labels are absent from feature extraction, Visual 1NN scoring, logit fusion, mass correction, and each prediction. MP-VTF performs no per-query adaptation or weight reselection in the reported runs. If the accumulated support cache contains N_S features of dimension d, Visual 1NN adds $O(N_S d)$ similarity work per query and $O(N_S d)$ cached-feature storage before class-wise maxima.

4 Experiments and Results

4.1 Experimental Setup

We evaluated CIFAR100, CUB-200-2011 (CUB200), and miniImageNet under the FSCIL protocols used by Bi-level Modality Calibration (BiMC) (Chen et al., 2025). Table 1 summarizes the class schedules, 5-shot increments, total sessions, and dataset-specific Masked Ensemble weights. Every inference path used random seeds 1, 2, and 3 following the BiMC support-sampling protocol.

We compared three matched inference paths: the BiMC baseline, Naive fusion, and Mass-Preserving Visual–Text Fusion (MP-VTF). All paths used the same frozen CLIP ViT-B/16 encoder, BiMC configuration, labeled support sets, test queries, random seeds, and official pre-generated category descriptions. Naive fusion and MP-VTF used fixed Text-CMNN/Visual 1NN weights of 0.9/0.1; their only difference was MP-VTF's query-wise log-sum-exp (LSE) correction. The visual weight was fixed before the nine reported runs and was not reselected within them (Section 3.5).

Avg denotes the mean all-class accuracy across sessions, and Final denotes the all-class accuracy at the final session of the complete fusion output. The base metric is the final-session accuracy over the base classes, and novel is the arithmetic mean of the final-session accuracies over the accumulated incremental class groups; H is their harmonic mean, $H = 2 \cdot \mathrm{base} \cdot \mathrm{novel}/(\mathrm{base} + \mathrm{novel})$. We report percentages as mean $\pm$ sample standard deviation across three seeds ($\mathrm{ddof} = 1$), and report between-method differences in percentage points (pp). Between-method differences are the mean of the three matched per-seed differences.

4.2 Novel-Group Mass Shift at the Fusion Interface

Under the fixed 0.9/0.1 Naive interface, the session-level mean novel-group probability shift was positive in all 78 incremental session–seed records (Figure 2; Table 2). This quantity is the within-record mean of the signed per-query shift δ^{Naive} (Section 3.3), where positive values indicate increased novel-group mass relative to Text-CMNN.

Across all 78 MP-VTF incremental session–seed records, the largest session-level mean absolute novel-group probability residual was 6×10^{-8} (Figure 2; Table 2) and the maximum absolute novel LSE residual was 4.8×10^{-7} (Table 2), verifying numerical satisfaction of the constraint.

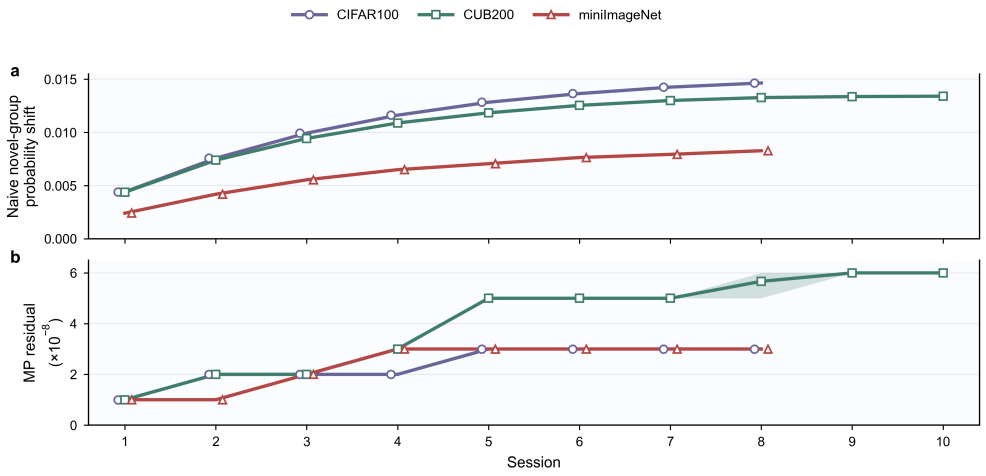

Figure 2: Per-session diagnostics of the fusion interface. (a) Mean signed novel-group probability shift of Naive fusion relative to Text-CMNN. (b) Mean absolute novel-group probability residual of MP-VTF, displayed after multiplication by 10^8. Each curve is the mean over three random seeds, and the shaded band spans their minimum and maximum. The largest seed spread at any session is 2.0×10^{-4} in (a) and 1.0×10^{-8} in (b). The x axis is the incremental session (1–8 for CIFAR100 and miniImageNet, 1–10 for CUB200). The panels use independent y axes.

Table 2: Fusion-interface diagnostics

Dataset	Naive shift	Positive records	Probability residual	LSE residual
CIFAR100	0.0043–0.0147	24/24	3.0×10^{-8}	2.4×10^{-7}
CUB200	0.0043–0.0134	30/30	6.0×10^{-8}	4.8×10^{-7}
miniImageNet	0.0024–0.0083	24/24	3.0×10^{-8}	2.4×10^{-7}

Naive shift is the min–max session-level mean signed probability shift, and positive records counts session–seed means above zero. The MP-VTF columns report the maximum session-level mean absolute novel-group probability residual and maximum absolute novel LSE residual. Probability and LSE residuals are not numerically interchangeable.

4.3 Endpoint Accuracy and Base–Novel Trade-offs

Under Naive fusion without the mass constraint, Avg and base accuracy were lower and novel accuracy was higher than in matched BiMC runs (Table 3). H decreased on CIFAR100 and CUB200 while increasing slightly on miniImageNet.

Relative to matched BiMC, MP-VTF changed dataset-level mean Avg by +0.023, +0.002, and +0.016 pp, and mean Final by +0.103, +0.013, and +0.020 pp on CIFAR100, CUB200, and miniImageNet, respectively (Table 3). Overall accuracy therefore remained close to BiMC.

MP-VTF increased novel accuracy in all three seeds on CIFAR100 and CUB200, and H increased in five of the six dataset–seed comparisons. The dataset-level means for novel accuracy and H were also higher than matched BiMC on all three benchmarks. Comparing Naive fusion with MP-VTF isolates the LSE correction at a fixed fused-logit input, whereas comparing MP-VTF with BiMC evaluates the complete fused interface (Table 3; Figure 4).

Figure 3 shows the longitudinal pattern: Naive fusion was below BiMC at every evaluated incremental session, whereas MP-VTF closely tracked BiMC across all three benchmarks.

Appendix A provides method-level context from reported and reproduced FSCIL results.

Table 3: Main FSCIL results

Dataset	Method	Avg	Final	base	novel	H
CIFAR100	BiMC	76.851 ± 0.175	73.223 ± 0.248	$\mathbf{77.763 \pm 0.225}$	66.417 ± 0.301	71.643 ± 0.264
CIFAR100	Naive fusion	74.532 ± 0.095	71.017 ± 0.442	67.703 ± 0.540	$\mathbf{75.983 \pm 0.319}$	71.604 ± 0.433
CIFAR100	MP-VTF (ours)	$\mathbf{76.874 \pm 0.177}$	$\mathbf{73.327 \pm 0.215}$	77.660 ± 0.125	66.825 ± 0.353	$\mathbf{71.836 \pm 0.256}$
CUB200	BiMC	75.760 ± 0.221	72.457 ± 0.404	$\mathbf{78.830 \pm 0.240}$	66.221 ± 0.912	71.975 ± 0.497
CUB200	Naive fusion	74.747 ± 0.159	71.730 ± 0.425	76.757 ± 0.416	$\mathbf{66.831 \pm 0.977}$	71.447 ± 0.501
CUB200	MP-VTF (ours)	$\mathbf{75.762 \pm 0.199}$	$\mathbf{72.470 \pm 0.419}$	78.740 ± 0.305	66.350 ± 0.976	$\mathbf{72.013 \pm 0.521}$
miniImageNet	BiMC	93.612 ± 0.053	92.357 ± 0.067	$\mathbf{93.057 \pm 0.227}$	91.308 ± 0.247	92.174 ± 0.059
miniImageNet	Naive fusion	93.412 ± 0.057	92.097 ± 0.042	91.327 ± 0.136	$\mathbf{93.250 \pm 0.180}$	92.278 ± 0.046
miniImageNet	MP-VTF (ours)	$\mathbf{93.628 \pm 0.045}$	$\mathbf{92.377 \pm 0.106}$	92.650 ± 0.280	91.967 ± 0.170	$\mathbf{92.307 \pm 0.066}$

Results are percentages reported as mean $\pm$ sample standard deviation across three random seeds. Avg is the mean all-class accuracy across sessions. Final, base, novel, and H are computed at the final session, where novel is the mean accuracy over the accumulated incremental class groups and H is the base–novel harmonic mean. Bold indicates the highest mean within each dataset and metric; boldface does not denote statistical significance. Each displayed method mean is rounded independently, so a between-method difference can differ by 0.001 pp from subtraction of the printed entries.

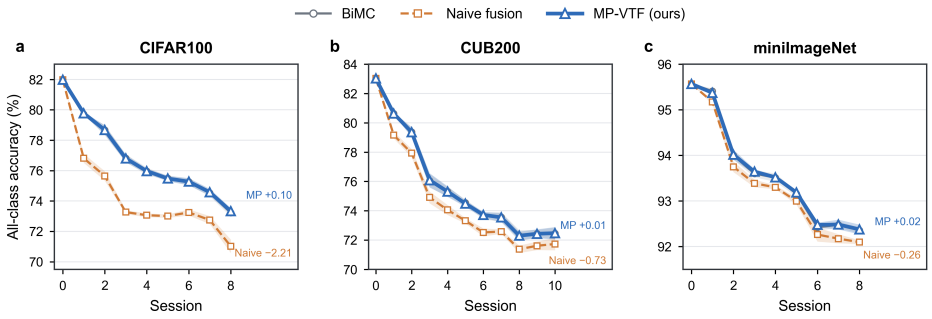

Figure 3: All-class accuracy across FSCIL sessions. Curves are means over three random seeds; shading is their sample standard deviation. Endpoint annotations give the final-session difference from matched BiMC in percentage points. Each dataset has an independent y axis.

5 Discussion

MP-VTF makes the Text-CMNN novel-group probability mass an explicit control variable at BiMC's globally normalized novel interface. Its query-wise correction removes the measured allocation shift while retaining the fused within-novel ordering. Across three benchmarks, overall accuracy remained close to BiMC and dataset-level mean novel accuracy was higher. Thus, support-image evidence changed distinctions among novel classes while the controlled Text-CMNN group mass stayed fixed.

Support-image scoring supplies class-specific evidence; MP-VTF adds an analytical LSE correction to control the group allocation that scoring can change. The correction adds no optimization stage and leaves BiMC's encoders, base logits, fused-prototype branch, ensemble weights, and decision rule unchanged.

The amount of visual evidence to add remains an empirical choice. Testing fixed visual weights and other frozen backbones under the same protocol would show how the base–novel balance changes while keeping group-level mass control intact.

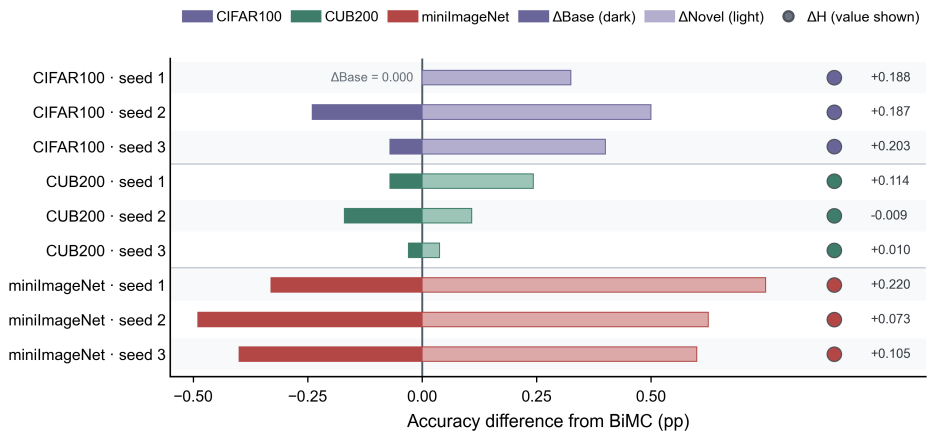

Figure 4: Matched-seed final-session changes from BiMC. Bars show ΔBase and ΔNovel; markers show ΔH for each dataset and seed.

6 Conclusion

At the evaluated BiMC novel interface, MP-VTF exactly restored the Text-CMNN novel-group probability mass while preserving the fused within-novel ordering. Across three benchmarks, it maintained overall accuracy close to BiMC and yielded higher dataset-level mean novel accuracy, providing a simple interface control for training-free FSCIL.

AI use statement

Generative AI tools assisted with drafting and editing parts of the manuscript, translating passages between Chinese and English, and revising Figure 1. The authors take responsibility for the final text, figure, and scientific claims.

References

Hongjoon Ahn, Jihwan Kwak, Subin Lim, Hyeonsu Bang, Hyojun Kim, and Taesup Moon. Ss-il: Separated softmax for incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 844–853. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00088.

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 (CVPR)*, pp. 9881–9890. IEEE, 2025. doi: 10.1109/CVPR52734.2025.00923.

Ali Cheraghian, Shafin Rahman, Sameera Ramasinghe, Pengfei Fang, Christian Simon, Lars Petersson, and Mehrtash Harandi. Synthesized feature based few-shot class-incremental learning on a mixture of subspaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8641–8650. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00854.

Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. MetaFSCIL: A meta-learning approach for few-shot class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14146–14155. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01377.

Marco D'Alessandro, Alberto Alonso, Enrique Calabrés, and Mikel Galar. Multimodal parameter-efficient few-shot class incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3385–3395. IEEE, 2023. doi: 10.1109/ICCVW60793.2023.00364.

Songlin Dong, Xiaopeng Hong, Xiaoyu Tao, Xinyuan Chang, Xing Wei, and Yihong Gong. Few-shot class-incremental learning via relation knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1255–1263, 2021. doi: 10.1609/aaai.v35i2.16213.

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*, volume 36, pp. 6582–6595. Curran Associates, Inc., 2023. doi: 10.52202/075280-0288.

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, 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL `https://proceedings.mlr.press/v70/guo17a.html`.

Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Constrained few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9047–9057. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00885.

Zitong Huang, Ze Chen, Zhixing Chen, Erjin Zhou, Xinxing Xu, Rick Siow Mong Goh, Yong Liu, Wangmeng Zuo, and Chunmei Feng. Learning prompt with distribution-based feature replay for few-shot class-incremental learning, 2024.

Haidong Kang, Ketong Qian, and Yi Lu. Breaking forgetting: Training-free few-shot class-incremental learning via conditional diffusion, 2025.

Eunju Lee, MiHyeon Kim, JuneHyoung Kwon, Yoonji Lee, JiHyun Kim, Soojin Jang, and YoungBin 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 (CVPR)*, pp. 29462–29471, 2026.

Yayuan Li, Jintao Guo, Lei Qi, Wenbin Li, and Yinghuan Shi. Text and image are mutually beneficial: Enhancing training-free few-shot classification with CLIP. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(5):5039–5047, 2025. doi: 10.1609/aaai.v39i5.32534.

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 (CVPR)*, pp. 23881–23890. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02254.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15691–15701. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01438.

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*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Oindrila Saha, Grant Van Horn, and Subhransu Maji. Improved zero-shot classification by adapting vlms with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17542–17552. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01661.

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 (CVPR)*, pp. 24183–24192. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02316.

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 (CVPR)*, pp. 12183–12192. IEEE, 2020. doi: 10.1109/CVPR42600.2020.01220.

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*, volume 36, pp. 15060–15076. Curran Associates, Inc., 2023. doi: 10.52202/075280-0661.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 374–382. IEEE, 2019. doi: 10.1109/CVPR.2019.00046.

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 (CVPR)*, pp. 12455–12464. IEEE, 2021. doi: 10.1109/CVPR46437.2021.01227.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of CLIP for few-shot classification. In *Computer Vision – ECCV 2022*, volume 13695 of *Lecture Notes in Computer Science*, pp. 493–510. Springer, 2022. doi: 10.1007/978-3-031-19833-5_29.

Hanbin Zhao, Yongjian Fu, Mintong Kang, Qi Tian, Fei Wu, and Xi Li. MgSvF: Multi-grained slow versus fast framework for few-shot class-incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1576–1588, 2024. doi: 10.1109/TPAMI.2021.3133897.

Linglan Zhao, Jing Lu, Yunlu Xu, Zhanzhan Cheng, Dashan Guo, Yi Niu, and Xiangzhong Fang. Few-shot class-incremental learning via class-aware bilateral distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11838–11847. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01139.

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 (CVPR)*, pp. 9046–9056. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00884.

Da-Wei Zhou, Han-Jia Ye, Liang Ma, Di Xie, Shiliang Pu, and De-Chuan Zhan. Few-shot class-incremental learning by sampling multi-phase tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12816–12831, 2023. doi: 10.1109/TPAMI.2022.3200865.

Kai Zhu, Yang Cao, Wei Zhai, Jie Cheng, and Zheng-Jun Zha. Self-promoted prototype refinement for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6797–6806. IEEE, 2021. doi: 10.1109/CVPR46437.2021.00673.

Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. Not all features matter: Enhancing few-shot CLIP with adaptive prior refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2605–2615. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00246.

A Additional Experimental Comparison

The method-level comparison below provides context across nominally aligned dataset protocols, but the reported and reproduced rows are not matched at the sample or seed level.

Table 4: Method-level FSCIL comparison under aligned dataset protocols

Method	Source	CIFAR100 Avg/Final	CUB200 Avg/Final	miniImageNet Avg/Final
CLIP Zero-Shot	Reported	71.33/68.93	59.23/55.73	88.76/86.15
Visual Prototype	Reported	67.49/61.03	73.04/68.80	88.55/86.09
TEEN	Reported	67.99/61.84	73.26/69.12	88.92/86.66
FeCAM	Reported	73.92/68.32	72.74/67.40	90.66/88.20
BiMC	Reproduced	76.85/73.22	**75.76**/72.46	93.61/92.36
MP-VTF	Reproduced	**76.87/73.33**	**75.76/72.47**	**93.63/92.38**
LP-DiF	Reported	–	–	93.76/91.68

External rows reproduce reported values from the comparison tables of BiMC (Chen et al., 2025), where CLIP Zero-Shot denotes the frozen zero-shot CLIP baseline (Radford et al., 2021), Visual Prototype the image-prototype baseline, TEEN follows (Wang et al., 2023), FeCAM follows (Goswami et al., 2023), and LP-DiF follows (Huang et al., 2024). Reproduced BiMC and MP-VTF rows are means over the matched runs in this study; all rows use CLIP ViT-B/16. Within the reproduced BiMC/MP-VTF pair, bold marks the higher displayed Avg and Final for each dataset; ties are both bolded. Boldface does not compare reported with reproduced rows or denote statistical significance.

B Additional Query-Level Diagnostic Analysis

B.1 Scope, Definitions, and Independence

This post-hoc diagnostic is restricted to CUB200 seed 2 and its ten incremental sessions. The source contains 44,674 aligned session-query rows. A prediction-change event is a row in which the BiMC baseline and MP-VTF predictions differ. Help denotes a changed prediction for which BiMC was incorrect and MP-VTF was correct; harm denotes the reverse. Query labels define help and harm only after prediction and never enter frozen inference.

Event-level records are not independent because the same query can recur across sessions. We therefore also deduplicate prediction changes by stable query ID, counting each query once across the ten sessions. The pooled AUROC analysis below retains repeated session-level events and is descriptive.

B.2 Prediction Changes and Branch Evidence

Across the ten incremental sessions, 345 prediction-change events corresponded to 144 distinct queries after stable-query-ID deduplication. Among novel-to-novel changes, 28 distinct queries were helped and 27 were harmed. At the event level, Visual 1NN selected the correct top-1 class in 42 of 58 novel-to-novel help events. Text-CMNN had selected the correct top-1 class in 53 of 69 novel-to-novel harm events. Visual 1NN could therefore both correct and overwrite the Text-CMNN novel ordering. Mass preservation alone does not determine whether query-specific reranking is semantically correct.

B.3 Diagnostic Signals and Pooled Discrimination

Of the 345 prediction-change events, 220 met the help/harm definitions used for this analysis, comprising 108 help and 112 harm events. The area under the receiver operating characteristic curve (AUROC) treats help as the positive class. For each signal, the label-adaptive best-direction summary is $\max(\mathrm{AUROC}, 1 - \mathrm{AUROC})$. It chooses direction on the same labeled events, is necessarily at least 0.5, and is reported only as a descriptive upper-bound summary. It does not resolve cross-session dependence.

The four query-label-free signals were the Text-CMNN novel margin, Visual 1NN novel margin, Visual 1NN top-1 score, and Text-Visual top-1 modality agreement. The agreement signal indicates whether the two branches select the same novel top-1 class. Their pooled best-direction AUROCs were 0.602, 0.570, 0.540, and 0.523, respectively.

B.4 Per-Session AUROC and Help/Harm Counts

Table S1 reports raw per-session AUROC. Table S2 reports the corresponding label-adaptive best-direction summaries. The abbreviated columns denote Text-CMNN novel margin (Text margin), Visual 1NN novel margin (Visual margin), Visual 1NN top-1 score (Visual score), and Text-Visual top-1 modality agreement (Agreement).

Table S1. Per-session raw AUROC for the four query-label-free diagnostic signals.

Session	help/harm	Text margin	Visual margin	Visual score	Agreement
1	2/5	0.5000	0.2000	0.6000	0.5500
2	6/7	0.6667	0.6190	0.3095	0.4524
3	8/9	0.5000	0.7639	0.4167	0.5208
4	8/10	0.4688	0.8375	0.5625	0.7000
5	7/11	0.5649	0.5974	0.2727	0.3896
6	14/8	0.5223	0.5848	0.4554	0.4732
7	11/13	0.6608	0.4860	0.6119	0.4371
8	12/11	0.6667	0.5152	0.5530	0.4053
9	23/19	0.5789	0.5767	0.6625	0.5252
10	17/19	0.7167	0.5960	0.6471	0.4505

Table S2. Per-session label-adaptive best-direction AUROC for the four query-label-free diagnostic signals. Direction is selected on the same labeled events, so values are descriptive upper-bound summaries rather than validation estimates.

Session	help/harm	Text margin	Visual margin	Visual score	Agreement
1	2/5	0.5000	0.8000	0.6000	0.5500
2	6/7	0.6667	0.6190	0.6905	0.5476
3	8/9	0.5000	0.7639	0.5833	0.5208
4	8/10	0.5312	0.8375	0.5625	0.7000
5	7/11	0.5649	0.5974	0.7273	0.6104
6	14/8	0.5223	0.5848	0.5446	0.5268
7	11/13	0.6608	0.5140	0.6119	0.5629
8	12/11	0.6667	0.5152	0.5530	0.5947
9	23/19	0.5789	0.5767	0.6625	0.5252
10	17/19	0.7167	0.5960	0.6471	0.5495

B.5 Interpretation Boundary

The per-session label-adaptive best-direction AUROC ranges were 0.50-0.72, 0.51-0.84, 0.54-0.73, and 0.52-0.70 for Text margin, Visual margin, Visual score, and Agreement. Raw direction varied: Agreement was below 0.5 in six sessions and Visual score in four. Help or harm counts ranged from 2 to 23, so these same-sample, direction-selected AUROCs are high-variance descriptive quantities. They do not justify an inference-time gate, establish general Visual 1NN unreliability, or generalize beyond this CUB200 seed-2 diagnostic.