GIP-GEO: GEOMETRIC SELF-EVALUATION FOR FEW-SHOT CLASS-INCREMENTAL LEARNING

Anonymous authors

ABSTRACT

Few-shot class-incremental learning requires learning novel classes from few labeled examples while retaining previously learned classes. Class-name-conditioned diffusion supports training-free adaptation, but its descriptions may not reflect the observed support images, and its generations are typically used without an explicit alignment check. GIP-Geo addresses both limitations. Grounded Iterative Prompting conditions the language model on support images and nearby base-class names, producing candidate descriptions of observed, discriminative attributes. Geometric self-evaluation then scores agreement among each candidate's text, real-support prototype, and generated prototype via alignment-regularized volume (ARV), which combines a kernel Gram determinant with pairwise dispersion to avoid a degenerate zero score. The same score selects one candidate and its fusion weight with real supports. On CIFAR-100, miniImageNet, and CUB-200, exhaustive GIP-Geo improves average accuracy over the controlled CD-FSCIL baseline without incremental network training; an optional score proxy reduces generation cost while CLIP and diffusion remain frozen.

1 INTRODUCTION

Few-shot class-incremental learning (FSCIL) introduces novel classes sequentially from only a few labeled examples while requiring a classifier to retain all previously learned classes. Freezing the representation prevents parameter drift, but it also makes each novel classifier weight depend heavily on a small support set. Conditional generation offers a complementary source of evidence: CD-FSCIL (Kang et al., 2025) generates class-conditioned exemplars and combines their representations with real supports without updating the representation during incremental sessions.

Conditional generation adds evidence, but two choices remain unresolved: *condition construction* and *evidence use*. Class names are broad; supports reveal appearances that distinguish novel classes from stored ones. Even a plausible description may not guide the generator or yield a prototype aligned with observed supports. The class-name-conditioned, fixed-fusion baseline leaves both decisions without support-grounded assessment.

We formulate diffusion-based FSCIL around two decisions: how to ground a generation condition and how to assess its output. Grounded Iterative Prompting (GIP) conditions the language model on support images and retrieved base-class names, producing candidate descriptions of observed, discriminative attributes. Geometric self-evaluation scores each text–support–generation path. GIP-Geo uses this score to select a candidate and fuse its prototype when a class enters the expanding classifier (Li et al., 2025).

Alignment-regularized volume (ARV) measures agreement among each candidate's text, real-support prototype, and generated prototype to rank candidates. A separate fusion objective scores the mixed prototype with a base-separation penalty to choose the real-support weight. Pairwise dispersion removes the determinant's zero-score ambiguity when exactly two inputs coincide. Exhaustive evaluation makes no network updates after base-session preparation; the reliability penalty and learned score proxy remain optional extensions.

This work makes three contributions:

- We introduce GIP to ground candidate descriptions in support images and retrieved base-class contrasts.

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.

- We use ARV to select candidates and fusion weights; pairwise dispersion removes the determinant's zero score when two inputs coincide.
- In controlled single-run comparisons, the correction-off core improves session-average accuracy over CD-FSCIL by 2.10–2.41 percentage points across three benchmarks. A separate correction-off evaluation shows higher average and final accuracy in all eight matched support-seed pairs; the reliability penalty and score proxy are evaluated only as optional extensions.

2 Related Work

Few-shot incremental learning. FSCIL methods preserve prior knowledge through topology or graph-based classifier evolution (Tao et al., 2020; Zhang et al., 2021), reserve or regularize representation space (Zhou et al., 2022; Zou et al., 2022; Yang et al., 2023; Song et al., 2023; Ahmed et al., 2024), and use rehearsal, replay, meta-learning, stronger representations, or weight-space ensembles (Rebuffi et al., 2017; Liu et al., 2022; Chi et al., 2022; Oh et al., 2024; Lee et al., 2025). A recent method freezes the visual encoder and learns class- and task-specific prototype offsets through per-session training (Jiang et al., 2026). GIP-Geo forms each novel prototype from support and generated evidence at class insertion, without incremental classifier tuning. We do not report a matched comparison with this per-session optimization method.

Vision-language and generative methods. Contrastive Language–Image Pre-training (CLIP) (Radford et al., 2021) provides the shared image–text space used in multimodal prompting (Khattak et al., 2023), feature replay (Huang et al., 2024), and frozen-prototype calibration (Chen et al., 2025; Lee et al., 2026). CD-FSCIL (Kang et al., 2025) generates class-conditioned exemplars for novel-class prototype fusion. Diffusion-FSCIL uses multi-scale Stable Diffusion features for incremental recognition and latent replay; DCS steers generation with classifier-derived rewards and updates its FSCIL classifier using generated images (Kim et al., 2025; Wu et al., 2025). GIP-Geo shares CD-FSCIL's generation-and-fusion setting and adds support-grounded triplet assessment for candidate selection and prototype fusion. CLIP and diffusion remain frozen during incremental sessions. Diffusion-FSCIL and DCS address feature replay or classifier updates. We do not report matched comparisons with either method.

Grounding, geometry, and reliability. VT-FSL (Li et al., 2025) is the closest conceptual predecessor. It combines support-conditioned four-stage prompting with kernelized-volume alignment of textual, support, and synthetic representations. GIP-Geo adapts triplet assessment to an expanding classifier, adds base-class contrasts, and uses the score for candidate selection and prototype fusion. Table 1 summarizes these design differences. VT-FSL evaluates episodic few-shot classification, whereas GIP-Geo maintains persistent prototypes as classes arrive. We do not report a direct transfer baseline, so the gain from adapting VT-FSL to FSCIL remains unmeasured. The real-support vector anchors the triplet in observations; pairwise measures such as CLIPScore (Hessel et al., 2021) compare two modalities. The supplementary logit penalty adjusts class scores and should not be confused with probability calibration (Guo et al., 2017).

3 Problem Formulation and Design Requirements

At incremental session t, the learner receives a K-shot support set for each new class, a frozen image–text representation, and the prototypes of previously seen classes. It must add one prototype per new class without using query labels or updating the representation. Conditional diffusion can provide additional evidence, but it introduces a path from a textual condition to generated images and then to a stored prototype. An error at any point in this path can affect all later queries assigned against that prototype.

Counterfactual: retaining multiple experts. Retaining a prototype for every candidate would postpone commitment to a single representation, allowing query-time aggregation or routing among candidates. This alternative increases stored representations and query comparisons as classes accumulate. We retain the standard one-prototype-per-class interface and make the decision at class

Table 1: Explicit differences between GIP-Geo and its closest predecessor, VT-FSL.

Aspect	VT-FSL	GIP-Geo
Task setting	Episodic few-shot classification	Class-incremental learning over an expanding classifier
Prompt context	Support images	Support images plus retrieved base-class names as contrasts
Assessment score	Kernelized volume of the text–support–synthetic triplet	Volume plus pairwise dispersion, removing the two-coincidence zero score
Use of geometry	Training-time cross-modal alignment; a large language model (LLM) filters generated images	Inference-time candidate scoring and fusion-weight search
Base-class handling	No persistent classifier	Base-name contrasts in prompting; base-separation penalty in fusion

introduction. The resulting problem is to construct plausible alternatives, assess them without query labels, and consolidate the selected evidence into one prototype.

Design requirements. This constraint yields three requirements. First, *condition construction* should reflect the observed supports and relevant base-class contrasts; a class name alone underdetermines the appearance of a few-shot episode. Second, *generation assessment* should evaluate the complete path. Text–support agreement checks the condition against observations, generated–support agreement checks whether synthesis remains anchored, and text–generation agreement checks the response to the condition. Each pair exposes a different source of mismatch. Third, *bounded consolidation* should select and fuse evidence without query labels or incremental network updates, with a cheaper approximation when exhaustive generation is unaffordable.

These requirements organize the method into two stages. GIP constructs support-grounded conditions with base-class contrasts; ARV evaluates each resulting triplet to select a path and construct one fused prototype. Classification correction and efficient inference extend this core sequence.

4 METHOD

Following the requirements in Section 3, Figure 1 summarizes two stages: GIP constructs support-grounded generation conditions, and ARV assesses the resulting evidence to select and fuse a prototype. We describe these stages in order, followed by the classification correction and optional efficiency extension.

4.1 SETTING

Let $\mathcal{C}_0$ be the base classes and $\mathcal{C}_t$ the novel classes introduced at session t. Each novel class c has a K-shot support set $\mathcal{S}_c$; queries are classified over $\bigcup_{i=0}^{t} \mathcal{C}_i$. We use frozen CLIP image and text encoders, with ϕ denoting the image encoder. The diffusion model is trained on base classes before incremental evaluation and is then frozen. Here, *training-free* refers to the exhaustive variant during incremental sessions, after base-session preparation. The optional proxy variant includes lightweight online training. Write $\mathrm{norm}(z) = z/\|z\|_2$ and form

$$p_c^{\mathrm{real}} = \mathrm{norm}\left(\frac{1}{K}\sum_{x\in\mathcal{S}_c} \mathrm{norm}(\phi(x))\right). \tag{1}$$

All prototypes and text embeddings below are normalized; the theoretical statements assume nonzero means.

4.2 GIP: GROUNDING THE GENERATION CONDITIONS

To address the mismatch between a class-name description and the observed examples, GIP supplies GPT-4o with the support images as well as the class name. It also retrieves the three nearest Session 0 class names in CLIP text space: these neighbors identify semantic contrasts to consider

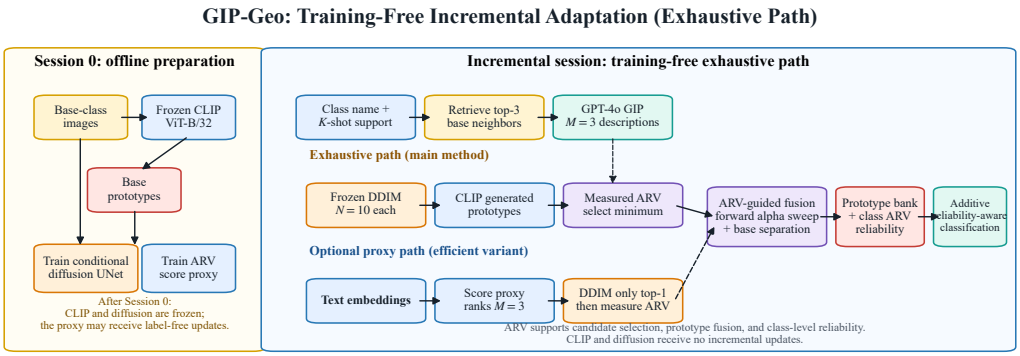

Figure 1: GIP-Geo uses supports and base-class context to construct $M = 3$ candidate conditions; ARV selects one path and fuses its generated prototype with real supports. The exhaustive core is training-free after Session 0. The class-level reliability correction is exploratory and disabled in the default pipeline and core results. The optional proxy ranks candidates before diffusion to reduce cost; only the proxy is trained online. CLIP ViT-B/32 and diffusion remain frozen during incremental sessions.

when describing a new class that must coexist with existing classes. A single call follows four stages: strategy, perception, refinement, and conclusion. Here, iterative prompting refers to refinement within this structured call. Refinement asks for visible attributes that distinguish the class from the retrieved neighbors. We sample $M = 3$ descriptions at temperatures $\{0.3, 0.5, 0.7\}$ to retain alternative conditions for evaluation. Their normalized CLIP text embeddings t_m condition a base-trained UNet through feature-wise linear modulation (FiLM) (Perez et al., 2018). For each candidate, the Denoising Diffusion Implicit Models (DDIM) sampler (Song et al., 2021) uses 50 deterministic steps to generate $N = 10$ exemplars, giving

$$p_{c,m}^{\text{gen}} = \text{norm}\left(\frac{1}{N}\sum_{i=1}^{N}\text{norm}(\phi(\hat{x}_{m,i}))\right). \tag{2}$$

DDIM uses $\eta = 0$, making generation deterministic conditional on the sampled initial noise. The next stage assesses the resulting text-to-generation paths before constructing the classifier. Implementation details and the complete algorithm appear in Appendix A.

4.3 ARV: Assessing and Using Generated Evidence

The second problem requires an assessment that can guide both which candidate to use and how to combine it with real supports. The text represents the intended condition, the generated prototype represents the generator's response, and the real prototype anchors both to observed examples. We use their agreement as a proxy for consistency, then apply it in two successive decisions: candidate selection and prototype fusion.

Constructing the assessment score. For three unit vectors $\mathcal{V} = \{v_1, v_2, v_3\}$, define the radial basis function (RBF) Gram matrix and its determinant volume by

$$G_{ij} = \exp\left(-\frac{\|v_i - v_j\|_2^2}{2\sigma^2}\right), \qquad V_{\text{det}}(\mathcal{V}) = \sqrt{\max(\det G, 0)}. \tag{3}$$

When two inputs coincide, G has duplicate rows and $V_{\text{det}} = 0$ regardless of the third input. We therefore use

$$D_k(\mathcal{V}) = \frac{1}{3}\sum_{i<j}(1 - G_{ij}), \qquad S_{\text{geo}}(\mathcal{V}) = V_{\text{det}}(\mathcal{V}) + \lambda_{\text{disp}} D_k(\mathcal{V}), \tag{4}$$

with $\sigma = 0.5$ and $\lambda_{\text{disp}} = 1$. For a positive regularizer, $S_{\text{geo}} = 0$ if and only if all three vectors coincide. Appendix B proves this property and gives conditional pairwise-distance and normalized-prototype bounds.

ARV combines two aspects of agreement. Dispersion measures mean pairwise mismatch; the determinant can distinguish some triplets with equal dispersion but different pairwise configurations (Appendix B). This provides a geometric criterion for selection and fusion whose practical value we test against simpler scores in Section 5.3.

Selecting a candidate. To make the choice depend on generated evidence, exhaustive selection evaluates the *unfused* triplet for every path:

$$S_{c,m} = S_{\text{geo}}(\{t_m, p_c^{\text{real}}, p_{c,m}^{\text{gen}}\}), \qquad m_c^* = \arg\min_m S_{c,m}. \tag{5}$$

Selection returns the path with the lowest measured inconsistency using only candidate text, generated prototypes, and labeled supports.

Fusing the selected evidence. Candidate selection determines which generation to use, but leaves its contribution to the final prototype unspecified. We use the same agreement criterion to choose a mixing weight for the selected path. Let $p_c(\alpha) = \text{norm}((1-\alpha)p_{c,m_c^*}^{\text{gen}} + \alpha p_c^{\text{real}})$. We sweep $\mathcal{A} = \{0, 0.05, \ldots, 1\}$ and select

$$\begin{aligned} \alpha_c^* &= \arg\min_{\alpha \in \mathcal{A}} \mathcal{L}_{\text{fuse}}(\alpha), \\ \mathcal{L}_{\text{fuse}}(\alpha) &= S_{\text{geo}}(\{p_c(\alpha), t_{m_c^*}, p_c^{\text{real}}\}) + \lambda \max\left(0, \max_{b \in \mathcal{C}_0} p_c(\alpha)^\top p_b - \delta_b\right), \end{aligned} \tag{6}$$

where $\lambda = 0.1$ and $\delta_b = 0.3$. The base term discourages excessive similarity to the well-supported Session 0 prototypes. It complements GIP's base-name contrasts at the feature level; competition among novel classes remains unconstrained. The forward sweep updates no network parameters. Here α is the real-support weight: $\alpha = 0$ retains only the generated prototype, and $\alpha = 1$ retains only the real prototype. At $\alpha = 1$, the determinant vanishes regardless of text–support agreement. Dispersion keeps the score positive when those inputs differ, while still allowing a support-only solution to minimize the objective. Candidate selection chooses the generated path; fusion sets its contribution alongside the observed supports.

4.4 Extensions: Classification Correction and Efficient Inference

Exploratory: reusing the assessment at classification. After prototype construction, the selected path's inconsistency could inform a class-level correction. We explore an additive penalty based on the selected *source* score $S_c^* = S_{c,m_c^*}$. The minimized fusion objective has a separate role: it chooses the mixing weight. Novel-class scores are min–max normalized within their introduction session to $\widetilde{S}_c \in [0,1]$, with 10^{-12} added to the denominator. A unit query q receives

$$\text{score}(q,c) = q^\top p_c(\alpha_c^*) - \lambda_{\text{conf}} \mathbb{I}[c \notin \mathcal{C}_0] \widetilde{S}_c. \tag{7}$$

For base classes, the stored base prototype replaces $p_c(\alpha_c^*)$, and no penalty is applied. We exclude this correction from the default pipeline. With unscaled cosine logits in $[-1,1]$, a large λ_{conf} can suppress a novel class whenever $\widetilde{S}_c > 2/\lambda_{\text{conf}}$. The observed outcomes also vary by support seed (Appendix E). We report the correction as an exploratory extension, including both its gains and failure modes.

Reducing the generation cost. Assessing every path requires diffusion for all three descriptions. As an optional efficiency extension, a residual multilayer perceptron (MLP) predicts source ARV from each text embedding and selects one description before diffusion. It is trained on base-class candidate sets with regression and ranking losses. The selected path's measured ARV provides the class reliability score and the target for five stochastic gradient descent (SGD) refinement steps, without query labels. This variant trains the proxy online; CLIP and diffusion remain frozen. The proxy approximates exhaustive selection, whose cost and accuracy serve as the reference. Architecture, losses, and ranking diagnostics are in Appendices A and F.

5 Experiments

We first report the correction-off core comparison, then isolate condition construction and evidence use. The final experiments examine reliability and efficiency extensions.

Table 2: Previously reported FSCIL results and a controlled CD-FSCIL/GIP-Geo comparison. Avg denotes the mean over sessions; Final denotes final-session accuracy. Earlier-method rows are literature results for benchmark context. The CD-FSCIL row is our controlled reproduction; only the final two rows form the matched single-run comparison. CIFAR-100 averages include source-reported Session-1 values with denominators under verification.

Method	Venue	CIFAR-100		miniImageNet		CUB-200	
		A_{avg}	Final	A_{avg}	Final	A_{avg}	Final
TOPIC (Tao et al., 2020)	CVPR 2020	64.23	52.78	61.34	49.18	68.42	62.90
CEC (Zhang et al., 2021)	CVPR 2021	67.85	56.86	64.52	52.76	71.36	65.83
FACT (Zhou et al., 2022)	CVPR 2022	69.41	58.61	66.18	54.63	73.54	68.15
CLOM (Zou et al., 2022)	NeurIPS 2022	70.92	60.31	67.85	56.50	75.12	69.62
CD-FSCIL (Kang et al., 2025)	arXiv 2025	73.56	63.28	70.43	59.41	77.85	72.62
GIP-Geo (exhaustive)	Ours	**75.66**	**66.72**	**72.61**	**62.61**	**80.26**	**75.33**
Δ		+2.10	+3.44	+2.18	+3.20	+2.41	+2.71

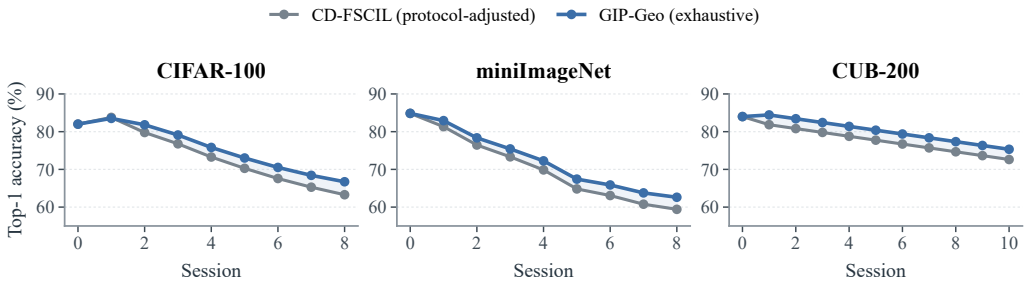

Figure 2: Single-run top-1 test-accuracy trajectories for the correction-off CD-FSCIL baseline and GIP-Geo on CIFAR-100, miniImageNet, and CUB-200, shown on a common vertical scale. Session averages and final-session values match the controlled rows in Table 2; the curves are not repeated-seed estimates. CIFAR-100 Session-1 values are retained pending verification of their query denominator.

5.1 Protocol and End-to-End Performance

CIFAR-100 and miniImageNet use 60 base classes followed by eight five-class sessions; CUB-200 uses 100 base classes followed by ten ten-class sessions. All novel classes have five supports. Let A_t be top-1 accuracy over all classes seen at session t. We report $A_{\text{avg}} = (T+1)^{-1} \sum_{t=0}^{T} A_t$ and final-session accuracy A_T. The controlled main-table results are single-run measurements: A_{avg} averages sessions within a run, not independent runs. Controlled CD-FSCIL/GIP-Geo comparisons use the same frozen CLIP ViT-B/32 encoder, diffusion backbone, class splits, and support seeds. Broader method rows provide benchmark context and should not be interpreted as isolating backbone or pretraining effects. Training and evaluation settings are detailed in Appendix A.

We first compare the complete core pipeline (without reliability penalty) with the controlled class-name-conditioned generation-and-fusion baseline. Table 2 reports the session-averaged and final-session accuracy. Figure 2 shows the session trajectories. In a separate correction-off evaluation, mean paired gains over eight matched support seeds ranged across datasets from 2.55 to 4.11 percentage points for session-average accuracy and from 4.28 to 7.10 points for final accuracy; all eight pairs were higher on each dataset (Appendix D.1). These descriptive results are separate from the single-run measurements in Table 2.

5.2 Grounding: The Contribution of Support-Conditioned Descriptions

On miniImageNet, class-name-only and support-conditioned variants without base contrasts report 70.81% and 72.13%, respectively (Table 3). The correction-off component reference is 72.39%, versus 72.61% in the main table; available records do not explain the gap. Replacement rows use

Table 3: Component measurements on miniImageNet. The full core row uses the correction-off setting and reports 72.39%, distinct from the 72.61% correction-off measurement in the main table; available records do not establish the source of this difference. Other component rows were measured under the earlier penalty-enabled setting. These values are descriptive, not matched estimates of component effects under the main-table run. Avg is computed over Sessions 0–8.

Configuration	S0	S1	S2	S3	S4	S5	S6	S7	S8	Avg
Full GIP-Geo (core, no penalty)	84.85	82.68	78.11	75.18	72.01	67.17	65.62	63.52	62.36	72.39
Grounded prompting										
w/o GIP (class-name-only text)	84.85	80.90	76.33	73.40	70.23	65.39	63.84	61.74	60.58	70.81
w/o base-class-aware prompting	84.85	82.39	77.82	74.89	71.72	66.88	65.33	63.23	62.07	72.13
Geometric self-evaluation										
w/o ARV selection (random M=3)	84.85	82.50	77.95	74.89	71.64	67.20	65.35	63.55	62.29	72.14
w/o geometric fusion ($\alpha = 0.5$ avg)	84.85	81.78	77.11	74.08	70.81	66.00	64.22	62.42	61.26	71.11
w/o base-class regularization	84.85	82.18	77.61	74.68	71.61	66.78	65.12	63.22	61.86	72.13
Efficiency extension										
Proxy (ranking + online refine)	84.85	82.82	78.15	75.62	72.45	67.91	65.86	63.16	62.50	72.60
Exploratory (not in default pipeline)										
+ reliability penalty (λ_{conf}=4.89)	84.85	82.63	78.36	75.63	72.62	67.22	65.67	63.88	62.51	72.91

an earlier penalty-enabled setting, and the class-name-only ablation changes prompting as well as image input. These contrasts therefore do not isolate grounding under the main policy. CIFAR-100 component records also lack matched-run metadata (Appendix D). We next test selection and fusion with grounded conditions.

5.3 Assessment: The Contribution of Candidate Selection and Fusion

Candidate selection. The reported random-selection measurement is 72.14% (Table 3). It uses the earlier penalty-enabled setting, while the current no-penalty full core measurement is 72.39%; the table does not provide a matched correction-off random-selection result. Thus, the selection effect under the main setting remains unmeasured by this ablation.

Prototype construction. The earlier penalty-enabled miniImageNet measurements report 71.11% with fixed $\alpha = 0.5$ and 72.13% without base-class regularization, alongside a 72.91% penalty-enabled full reference (Table 3). These measurements vary the mixing weight and base-class competition separately, but do not estimate those effects relative to the current correction-off pipeline (Figure 3(b)).

Determinant regularization. In one CIFAR-100 run, adding dispersion raised fusion accuracy from 74.91% to 75.88% (Figure 3(a)). Across five support-seed replicates, ARV reduced $\alpha = 1$ selections from 46.5%, 44.0%, and 50.4% to 1.5%, 1.5%, and 1.0% on CIFAR-100, miniImageNet, and CUB-200, respectively (Figure 6). This is consistent with removing the endpoint's automatic zero score. The accuracy difference is a single-run estimate; the replicated diagnostic measures selection frequency, not accuracy gains.

Comparison with simpler agreement scores. The eight-replicate fusion comparison in Figure 4 shows a positive ARV-minus-determinant interval; the dispersion and cosine intervals include zero. Matched-candidate controls in Appendix C show similar accuracy for ARV and simpler scores, with small dataset-specific reversals. Dispersion-only selection reaches 75.96% on CIFAR-100, and cosine-disagreement selection reaches 72.66% on miniImageNet, compared with 75.88% and 72.61% for ARV. The results support two conclusions: dispersion removes the determinant's endpoint degeneracy, and ARV exceeds determinant-only scoring in the reported single-run comparison. They do not establish an accuracy advantage over dispersion-only or cosine scores. ARV also supplies one criterion for candidate selection, fusion, and the reliability analysis. End-to-end performance is reported in Table 2.

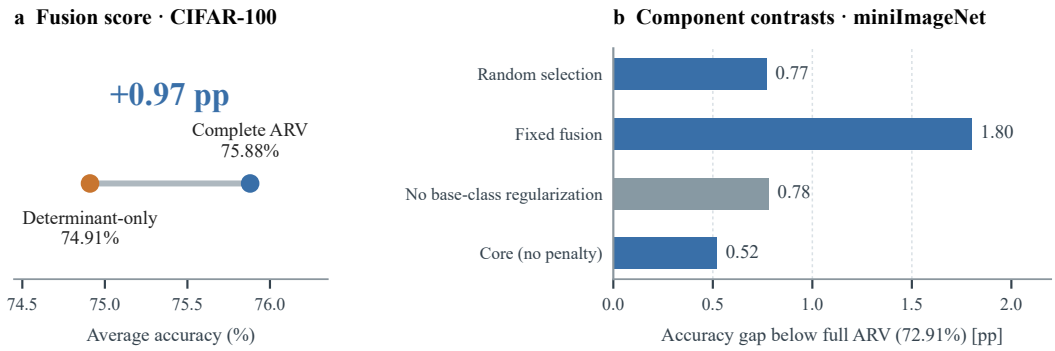

Figure 3: Assessing how the score is used. (a) Reported single-run CIFAR-100 determinant-only and ARV fusion comparison. (b) Descriptive miniImageNet component contrasts relative to the 72.91% full-method result with reliability penalty; replacement measurements come from an earlier penalty-enabled run and are not matched effects under the main correction-off setting. The panels address separate decisions; their effects are not additive, and no uncertainty intervals are shown.

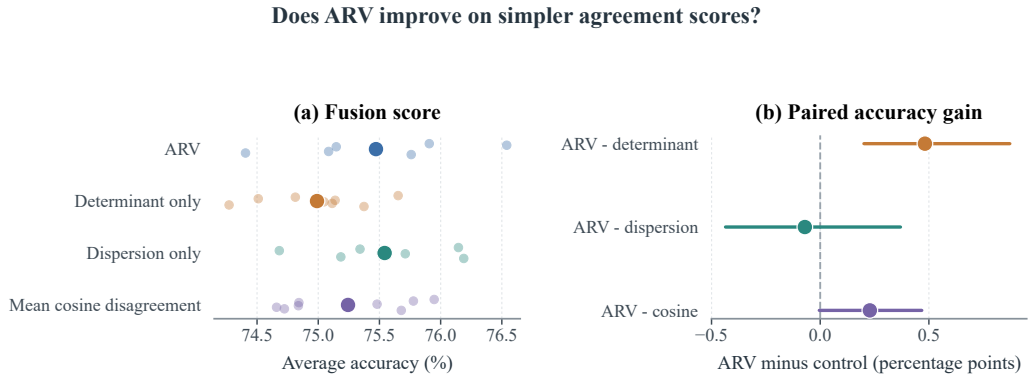

Figure 4: Fusion-score comparison on CIFAR-100 over eight independent replicates. (a) Small points show replicate-level average accuracy; large points show their arithmetic means. (b) Paired ARV-minus-control differences, with 95% percentile bootstrap confidence intervals from 20,000 resamples of the eight paired differences. Positive values favor ARV; intervals crossing zero leave the direction unresolved.

5.4 Exploratory Extensions: Reliability Penalty and Generation Cost

Reliability penalty at classification. As an exploratory extension, we test whether ARV can inform classification. We select $\lambda_{\text{conf}} = 4.89$ once on CIFAR-100 base-class validation episodes and hold it fixed across benchmarks and support seeds (Appendix A). In one CIFAR-100 run, the correction increases novel-class accuracy from 48.0% to 50.2% and base accuracy from 74.6% to 75.1%. On a different support seed (seed 101), the same correction lowers novel accuracy from 65.92% to 18.33%, with zero correct predictions for 17 of 40 novel classes. This instability arises because the additive penalty acts as hard class deletion when $\widetilde{S}_c > 2/\lambda_{\text{conf}} \approx 0.409$. We therefore exclude this correction from the core pipeline and main results. Appendix E reports both cases in detail.

Approximating exhaustive assessment efficiently. CIFAR-100 averages 75.87% with the proxy and 75.88% with exhaustive ARV. On miniImageNet, the corresponding values are 72.60% and 72.61%. On CUB-200, both modes make 300 LLM calls, and the proxy reduces generated images from 3,000 to 1,000. Reported runtime falls from 3.0× to 1.3× the CD-FSCIL reference. The available records omit hardware, absolute elapsed times, and timer boundaries; these ratios describe the reported runs and do not provide reproducible latency estimates (Appendix F). The appendix also

reports ranking fidelity and resource accounting. The proxy lowers generation cost by approximating candidate selection, with lightweight online training.

6 Discussion

Mechanism and claim scope. GIP turns support images and base-class contrasts into candidate conditions. ARV scores each text–support–generation triplet, then guides candidate selection and prototype fusion. Pairwise dispersion keeps the score positive when exactly two inputs coincide. The component comparisons are single-run measurements; their stability across seeds remains untested. Dispersion-only and cosine scores remain competitive, so the experiments do not establish a unique accuracy benefit from the determinant term. The joint-stage comparison gives positive point estimates for both stages and no observed positive interaction (Appendix C.3). The separate one-component ablation gains should not be summed.

Relation to prior work. GIP-Geo adapts ideas from VT-FSL to an expanding classifier. Base-class contrasts guide description generation, and the score controls candidate selection and prototype fusion. Our experiments compare the method with CD-FSCIL and test its components. They do not include a matched-budget transfer of VT-FSL, so the gain from adapting VT-FSL to this setting remains unmeasured.

Efficiency. Exhaustive assessment uses more computation when each class is introduced, while the classifier stores one prototype per class. The proxy ranks candidates from text embeddings and uses lightweight online updates, reducing generation cost. Both variants incur this cost at class introduction; later queries use the same prototype-based classifier.

Robustness and failure modes. Both stages depend on representative supports. Misleading examples or systematically biased embeddings may produce a low score even for poor evidence. ARV measures agreement in embedding space; it cannot verify whether a description states a true visual attribute, and every candidate may be poor. We have no matched support-only or no-diffusion baseline, so the experiments do not isolate the contribution of generated images. Base-name retrieval and the separation term address contrasts with Session 0 classes. Competition among novel classes remains as the classifier expands. For $K = 5$ at 95% confidence, the normalized-prototype bound's perturbation term is at least 3.08. The bound therefore reaches the unit-sphere diameter of 2 and cannot certify candidate ordering (Appendix B). The eight support-seed pairs do not isolate sensitivity to diffusion noise, and session accuracies are correlated.

Reliability penalty. Only two runs show different outcomes across support sets, so transfer of the fixed coefficient remains unestablished. The correction remains outside the core method. Any bounded, logit-scale-aware alternative needs paired support-seed evaluation before broader performance claims.

7 Conclusion

GIP-Geo combines support-grounded condition generation with assessment of the resulting evidence. GIP produces candidate descriptions from support images and base-class contrasts; ARV scores each generated path and guides candidate selection and prototype fusion. The benchmark results support the combined pipeline, while component comparisons provide narrower evidence about individual decisions. The reliability penalty remains exploratory because its effects vary across support seeds. The optional proxy reduces generation cost and uses lightweight online training. The exhaustive method makes no incremental network updates.

AI USE STATEMENT

OpenAI GPT-4o was used within the method to generate support-grounded candidate descriptions. OpenAI Codex assisted with manuscript revision; literature and citation checking; feedback on experimental design and controls; interpretation of reported results; implementation of experiment-control code; visualization and formatting of existing reported measurements; and formulation and revision of mathematical arguments. The authors manually reviewed and verified the AI-assisted literature and citation checks, experiment-control code, figures, and mathematical arguments. The authors take responsibility for the final content of the submission, including its text, claims, code, and results.

REFERENCES

Noor Ahmed, Anna Kukleva, and Bernt Schiele. Orco: Towards better generalization via orthogonality and contrast for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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, 2025.

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)*, 2022.

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*, pp. 1321–1330, 2017.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

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. *arXiv preprint arXiv:2401.01598*, 2024.

Shengqin Jiang, Xiaoran Feng, Yuankai Qi, Haokui Zhang, Renlong Hang, Qingshan Liu, Lina Yao, Quan Z. Sheng, and Ming-Hsuan Yang. Unlocking prototype potential: An efficient tuning framework for few-shot class-incremental learning, 2026. URL https://arxiv.org/abs/2602.05271.

Haidong Kang, Ketong Qian, and Yi Lu. Breaking forgetting: Training-free few-shot class-incremental learning via conditional diffusion. *arXiv preprint arXiv:2511.18516*, 2025. URL https://arxiv.org/abs/2511.18516.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19113–19122, 2023.

Junsu Kim, Yunhoe Ku, Dongyoon Han, and Seungryul Baek. Beyond synthetic replays: Turning diffusion features into few-shot class-incremental learning knowledge, 2025. URL https://arxiv.org/abs/2503.23402.

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 (CVPR)*, 2026.

Juntae Lee, Munawar Hayat, and Sungrack Yun. Tripartite weight-space ensemble for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15329–15338, 2025.

Wenhao Li, Qiangchang Wang, Xianjing Meng, Zhibin Wu, and Yilong Yin. VT-FSL: Bridging vision and text with LLMs for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a52969d383a4588c8a3857017a0195f8-Abstract-Conference.html.

Huan Liu, Li Gu, Zhixiang Chi, Yang Wang, Yuanhao Yu, Jun Chen, and Jin Tang. Few-shot class-incremental learning via entropy-regularized data-free replay. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics, 1989*, pp. 148–188. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, 2021.

Junghun Oh, Sungyong Baik, and Kyoung Mu Lee. Closer: Towards better representation learning for few-shot class-incremental learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

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. *arXiv preprint arXiv:2103.00020*, 2021.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

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)*, 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 (CVPR)*, 2020.

Ruitao Wu, Yifan Zhao, Guangyao Chen, and Jia Li. Diffusion-classifier synergy: Reward-aligned learning via mutual boosting loop for FSCIL. In *Advances in Neural Information Processing Systems 38*, pp. 71590–71617, 2025. doi: 10.52202/085713-2153. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/5cc053392052b3ee4c2b912f4da62ab5-Abstract-Conference.html.

Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip H. S. Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In *International Conference on Learning Representations (ICLR)*, 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 (CVPR)*, 2022.

Yixiong Zou, Shanghang Zhang, Yuhua Li, and Ruixuan Li. Margin-based few-shot class-incremental learning with class-level overfitting mitigation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ae817e85f71ef86d5c9566598e185b89-Abstract-Conference.html.

A Method Walkthrough and Implementation

This appendix connects the two problems to the method decisions, then provides the implementation and algorithm. Subsequent appendices explain ARV (Appendix B), specify controlled evaluations (Appendix C), report accuracy results (Appendix D), and examine reliability and efficiency (Appendices E and F).

A.1 How the Two Stages Address the Two Problems

From observed attributes to a classification decision. For each novel class, the support set supplies the visual reference for both stages. GIP uses this reference to constrain the generation condition; ARV uses it again to assess the generator's response. The following sequence identifies what information enters each decision and which failure it addresses.

Stage	Information and decision
Class-name condition	The class name supplies category semantics, but does not specify which appearances are represented by the current supports or which visible attributes distinguish them from base classes.
Grounded condition	GIP supplies support images to constrain attributes such as color, shape, texture, and parts. Retrieved base-class names direct refinement toward visible distinctions.
Generation assessment	Different descriptions or diffusion realizations may retain, omit, or alter these attributes. ARV compares each text–support–generation triplet in CLIP space; it does not explicitly detect named attributes.
Candidate selection	The minimum source ARV selects one complete path. A fluent description alone does not determine the choice: its generated prototype also enters the score.
Prototype fusion	The grid search sets the selected generation's contribution relative to the supports, subject to the base-prototype penalty. The resulting prototype enters the expanding classifier.

The sequence separates two possible failures: an underspecified condition and an inconsistent response to a grounded condition. GIP acts before generation; ARV acts after generation. Omitting grounding leaves candidate construction dependent on class-name semantics. Omitting assessment leaves no consistency-based choice among generated candidates or their fusion weights. Agreement remains an embedding-space proxy, so a visually incorrect but mutually consistent triplet can still score well.

A.2 Implementation and Complete Algorithm

Figure 5 summarizes one incremental session, and Algorithm 1 gives the complete procedure.

Four-stage prompt. The strategy stage relates the class name to prior knowledge; perception records colors, shapes, textures, parts, and contexts shared by the supports; refinement removes unsupported or non-discriminative attributes; and conclusion produces the diffusion description. Refinement includes the instruction: *"Distinguish {class_name} from {top_1, top_2, top_3}; remove shared features and retain the most discriminative visible attributes."* The three names correspond to the nearest Session 0 class-name embeddings under CLIP cosine similarity.

Current source prompt and image input. For reproducibility, we include the currently available prompt template below. It documents the implementation source, but does not establish that every archived run used this exact text: historical API request bodies and a dated GPT-4o snapshot were not retained. The system message is *You are a visual recognition expert. Generate precise, visually grounded descriptions of object classes based on class names and support images. Use the structured reasoning stages provided.* Each support tensor is converted from $[-1, 1]$ to an 8-bit PNG and appended to the user message in support order as an image input with `detail=low`. The text portion is:

```
[Optional prefix when base-class neighbors are available]
Critically, distinguish {class_name} from these related base classes:
{neighbor_1}, {neighbor_2}, {neighbor_3}. Remove features shared with
{neighbor_1}, {neighbor_2}, {neighbor_3}, then emphasize the visual
evidence
that is discriminative for {class_name}.

You are a visual recognition expert. Your task is to generate a precise,
visually grounded description of the class "{class_name}".

<STRATEGY>
The class name "{class_name}" may refer to a specific object category.
First, I will interpret the class name and recall general knowledge about
it.
Then I will examine the provided support images to extract the shared
visual
features that define this category. Finally, I will cross-validate these
observations and produce a concise, detailed description that accurately
reflects both the semantic meaning of the class name and the actual
visual
patterns present in the images.
</STRATEGY>

Below are {num_shots} support images of this class (encoded as visual
context).

Support image analysis for class "{class_name}":
The images show [brief summary of what the support images depict].

<PERCEPTION>
After examining the {num_shots} support images of "{class_name}", I
observe
the following consistent visual patterns across all samples:
- Color: [describe the dominant and secondary colors, color patterns]
- Shape: [describe the overall shape, proportions, silhouette]
- Texture: [describe surface texture, patterns, material appearance]
- Key parts: [list distinctive parts/features that are consistently
present]
- Background/context: [note any consistent environmental cues]
These visual attributes are present in every support image, confirming
they
are defining characteristics of {class_name}.
</PERCEPTION>

<REFINEMENT>
```

```
Let me verify these observations against my general knowledge of
"{class_name}" and cross-reference with the support images:
1. The color patterns I observed [confirm/refine; e.g., "are consistent
with
   the typical coloration of this category"].
2. The shape characteristics [match the expected morphology / show some
   intra-class variation in...].
3. The textural features [are distinctive / may be shared with similar
   categories, so I will emphasize more unique aspects].
4. The key parts identified [are indeed diagnostic features that reliably
   distinguish this class from visually similar ones].
The most robust and discriminative visual features for "{class_name}"
are:
</REFINEMENT>

<CONCLUSION>
Based on the cross-referenced analysis above, the class "{class_name}"
can
be visually described as follows:
[concise, precise visual description that combines class-level semantics
 with the specific visual patterns observed in the support images.
 Include: shape, color, texture, distinctive parts, typical setting]
This description is grounded in both the semantic meaning of the class
name and the actual visual evidence from the {num_shots} support samples.
</CONCLUSION>
```

Diffusion conditioning. For a normalized text embedding t, the conditional UNet applies FiLM as

$$\mathrm{FiLM}(h_i \mid \gamma, \beta) = \gamma_i \odot h_i + \beta_i, \qquad [\gamma, \beta] = \mathrm{MLP}(t). \tag{8}$$

It is trained on base images with the standard denoising loss

$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}\left[\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{a}_s}x_0 + \sqrt{1-\bar{a}_s}\,\epsilon, s, t)\right\|_2^2\right], \tag{9}$$

using a cosine schedule (Nichol & Dhariwal, 2021) with $T = 1000$. Here $s \in \{1, \ldots, T\}$ indexes the diffusion timestep and $\bar{a}_s$ is the noise-schedule coefficient; they are distinct from the incremental session index t and the fusion weight α used elsewhere. In the available implementation, the cosine schedule uses offset 0.008; training images are resized to 64×64 and mapped to $[-1, 1]$, and the text condition is "a photo of a {class_name}."

Encoder and diffusion architecture. The available implementation uses an image-space UNet with three encoder stages of 128, 256, and 256 channels, a 256-channel bottleneck, and decoder stages of 128, 64, and 64 channels. Its residual convolution blocks use two 3×3 convolutions with GroupNorm(8) and SiLU; bilinear upsampling and skip concatenation connect the decoder. Text-conditioned FiLM applies channel-wise scale and shift, while timestep embeddings enter through a separate MLP. The frozen CLIP wrapper loads ViT-B/32, resizes feature-extraction images bilinearly to 224×224, maps them from $[-1, 1]$ to $[0, 1]$, and normalizes channels with means $(0.48145466, 0.4578275, 0.40821073)$ and standard deviations $(0.26862954, 0.26130258, 0.27577711)$. Text is tokenized with truncation enabled. These details describe the available implementation; they do not establish a code or checkpoint digest for each archived benchmark run.

Proxy architecture and objective. The proxy uses a 512–256 input projection with LayerNorm, ReLU, and dropout 0.1, followed by a two-layer 256-dimensional residual block. A dual-output head predicts $[z_s, \log v]$, with score $\hat{s} = \mathrm{softplus}(z_s)$. Log-variance is clipped to $[-10, 10]$ before evaluating the Gaussian loss. With n_b measured base-class candidate scores S_i, training minimizes

$$\mathcal{L}_{\mathrm{proxy}} = \frac{1}{n_b}\sum_i \left[\frac{(S_i - \hat{s}_i)^2}{2v_i} + \frac{1}{2}\log v_i\right] + \lambda_{\mathrm{rank}}\mathcal{L}_{\mathrm{rank}}. \tag{10}$$

Each AdamW step uses the full set of $3|\mathcal{C}_0|$ examples from base-class few-shot episodes, with weight decay 10^{-5} and gradient-norm clipping at 1. Online refinement uses five SGD steps at learning rate 5×10^{-5} on each newly measured text–score pair. Because this update contains only one candidate, it uses the Gaussian term alone. Candidate ranking uses the predicted score in evaluation mode;

the variance output is used only during training. Online refinement updates all proxy parameters, including the log-variance head, so minimizing the Gaussian term on a single pair can in principle drive the predicted variance toward its clipping bound. The log-variance clip, the small learning rate, and the five-step budget limit this drift, and selection never consumes the variance output, so such drift does not affect candidate ranking. Freezing the variance head during online refinement is a straightforward additional safeguard.

Sequential proxy state. The current runner follows the loader order specified below and trains a fresh proxy at Session 0. In proxy mode, selection precedes online refinement, and updated weights carry forward across classes and sessions. The archived ranking audit lacks a run-specific class-order record and initial proxy checkpoint, so its exact sequence cannot be reconstructed from the report alone.

Training and evaluation configuration. The conditional diffusion UNet operates on 64×64 images with base channels 64 and time-embedding dimension 256. It is trained for 200 epochs on base classes using AdamW with learning rate 10^{-4} and batch size 64. DDIM uses $\eta = 0$ and 50 steps; each call draws a fresh Gaussian initial latent, making its updates deterministic conditional on that draw. We use $N = 10$ exemplars per evaluated description: 30 images per novel class for exhaustive evaluation, 10 for the proxy. GIP uses GPT-4o with a 512-token output limit and $M = 3$ candidates at temperatures $\{0.3, 0.5, 0.7\}$. The proxy is trained for 200 epochs at learning rate 10^{-3} with ranking-loss weight $\lambda_{\text{rank}} = 0.5$. Encoder and ARV settings follow Section 4. The reliability coefficient $\lambda_{\text{conf}} = 4.89$ is selected on a validation set. Base-separation parameters are selected once on base-class validation episodes and fixed for all incremental sessions. The available implementation uses the API alias `gpt-4o`; archived benchmark records do not identify a dated model snapshot. We therefore cannot verify the exact GPT-4o version used in each historical run.

Dataset order and support sampling. The available loader assigns CIFAR-100 and miniImageNet class IDs 0–59 to base training and IDs 60–99 to novel sessions in ascending ID order, five classes per session. For CUB-200 it uses IDs 0–99 as base classes and 100–199 as novel classes, ten per session. For a class with ID c, support indices are sampled with a PyTorch generator seeded by $s + 1009c$, where s is the runner's `--seed`. This describes the available loader; the archived single-run main-table records do not preserve the actual `--seed` invocation or bind each result to a checkpoint digest.

Ranking-aware training. Standard regression treats samples independently, but the proxy must identify the best candidate among M. For each class c, let its candidates have measured scores $\{S_{c,m}\}_{m=1}^{M}$ and predicted scores $\{\hat{s}_{c,m}\}_{m=1}^{M}$. For every ordered pair (i, j) satisfying $S_{c,i} < S_{c,j}$, we use the hinge loss

$$\mathcal{L}_{\text{rank}} = \frac{1}{\sum_c |\mathcal{P}_c|} \sum_c \sum_{(i,j)\in\mathcal{P}_c} \max\big(0,\ \hat{s}_{c,i} - \hat{s}_{c,j} + \epsilon\big), \tag{11}$$

where $\mathcal{P}_c = \{(i, j) : S_{c,i} < S_{c,j}\}$ and $\epsilon = 10^{-4}$. The loss averages over all valid within-class pairs; equal-score pairs are excluded. If no valid pair exists, $\mathcal{L}_{\text{rank}} = 0$. This term trains the ordering used for candidate selection.

Selection and fusion conventions. Exhaustive selection compares source-triplet ARV before fusion. Fusion evaluates the 21 weights and retains the largest real-support weight on exact ties. The reliability penalty uses the selected source score; the minimized fusion objective only selects the mixing weight. If all novel classes in a session have equal source scores, min–max normalization gives zero penalties for that session. Stored session-normalized scores remain fixed as later sessions arrive.

Reliability-coefficient selection. The CIFAR-100 validation protocol uses 20 pseudo-incremental episodes drawn only from the 60 base classes. Each episode treats five classes as pseudo-novel and the remainder as pseudo-base. A seeded, class-stratified 20% holdout supplies queries; the remaining samples form pseudo-base prototypes or provide five supports per pseudo-novel class. Query samples enter neither prototypes nor support sets. Real incremental classes and benchmark test images are excluded. We search the inclusive grid $[0, 6]$ with step 0.01 (601 candidates). The

objective is the mean, across episodes, of the harmonic mean of pseudo-base and pseudo-novel query accuracy. Coefficients whose objective values differ from the maximum by at most 10^{-12} are tied; the smallest tied coefficient is selected. This procedure selects $\lambda_{\text{conf}} = 4.89$. The split seed is 2027, with episode seed $2027 + 1009e$ for episode index $e = 0, \ldots, 19$. The coefficient is selected once on CIFAR-100 base-class validation episodes and applied unchanged to other benchmarks and support seeds; Appendix E shows that this transfer is unreliable, which is one reason the correction is excluded from the default pipeline.

Algorithm 1 GIP-Geo for one incremental session

Input: Novel classes $\mathcal{C}_t$, supports $\{\mathcal{S}_c\}$, base prototypes $\{p_b\}$, mode $\in \{\textsc{Exhaustive}, \textsc{Proxy}\}$; optional proxy $\mathcal{V}$
Output: Updated prototype bank with reliability scores

for each novel class $c \in \mathcal{C}_t$ **do**
  $\mathcal{B}_c \leftarrow \text{TopKBase}(c, \mathcal{C}_0, 3)$ *top-3 semantic neighbors*
  $p_c^{\text{real}} \leftarrow \text{norm}(\frac{1}{K} \sum_{x \in \mathcal{S}_c} \text{norm}(\phi(x)))$
  for temperature $\tau \in \{0.3, 0.5, 0.7\}$ **do**
    $d_\tau \leftarrow \text{MLLM}(\text{FourStagePrompt}(c, \mathcal{B}_c), \mathcal{S}_c; \tau)$ *GIP*
    $t_\tau \leftarrow \text{norm}(\text{CLIP}_{\text{text}}(d_\tau))$
  end for
  if mode $= \textsc{Proxy}$ **then**
    $\tau^* \leftarrow \arg\min_\tau \mathcal{V}(t_\tau); \mathcal{I} \leftarrow \{\tau^*\}$
  else
    $\mathcal{I} \leftarrow \{0.3, 0.5, 0.7\}$
  end if
  for $\tau \in \mathcal{I}$ **do**
    $\{\hat{x}_{\tau,i}\}_{i=1}^N \leftarrow \text{DDIM}(\epsilon_\theta, t_\tau, N)$
    $p_{c,\tau}^{\text{gen}} \leftarrow \text{norm}(\frac{1}{N} \sum_i \text{norm}(\phi(\hat{x}_{\tau,i})))$
    $S_{c,\tau} \leftarrow S_{\text{geo}}(\{t_\tau, p_c^{\text{real}}, p_{c,\tau}^{\text{gen}}\})$
  end for
  if mode $= \textsc{Exhaustive}$ **then**
    $\tau^* \leftarrow \arg\min_{\tau \in \mathcal{I}} S_{c,\tau}$
  end if
  $p_c \leftarrow \text{Geo-Fuse}(p_{c,\tau^*}^{\text{gen}}, p_c^{\text{real}}, t_{\tau^*}, \{p_b\})$
  $S_c^* \leftarrow S_{c,\tau^*}$; add (p_c, S_c^*) to the prototype bank
  if mode $= \textsc{Proxy}$ **then**
    Refine $\mathcal{V}$ on (t_{τ^*}, S_c^*) for five steps
  end if
end for
Normalize $\{S_c^* : c \in \mathcal{C}_t\}$ within this session as in Section 4.4
return updated prototype bank

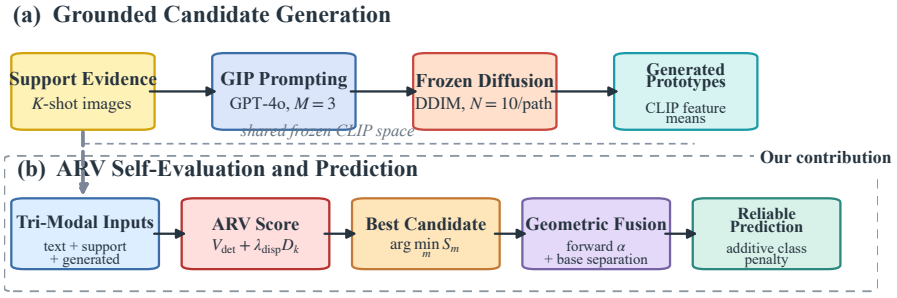

Figure 5: Overview of GIP-Geo in one incremental session. (a) Support-grounded GIP produces M candidate descriptions, and frozen DDIM maps each candidate to generated prototypes. (b) ARV jointly compares text, support, and generated embeddings, selects the best candidate, guides prototype fusion, and supplies class reliability. The efficient proxy path is detailed in Figure 1.

B ARV Design Rationale and Theoretical Properties

B.1 Why Combine Volume and Dispersion?

Write the off-diagonal RBF similarities as $a = k(t, r)$, $b = k(t, g)$, and $c = k(r, g)$. Then

$$D_k = 1 - \frac{a+b+c}{3}, \qquad V_{\text{det}} = \sqrt{1 + 2abc - a^2 - b^2 - c^2}. \tag{12}$$

Dispersion fixes the sum of pairwise similarities; volume also depends on their product and squared magnitudes. Consider the following constructed configurations at $\sigma = 0.5$:

Configuration (a, b, c)	D_k	V_{det}	ARV ($\lambda_{\text{disp}} = 1$)
$(1/3, 1/3, 1/3)$	$2/3$	$\sqrt{20/27}$	1.52733
$(1/2, 1/4, 1/4)$	$2/3$	$\sqrt{11/16}$	1.49582

Both are realizable by three unit vectors: the corresponding Gram matrices in the original feature space have diagonal one and off-diagonal entries $1 + \sigma^2 \log k$, and are positive definite. Thus ARV can distinguish triplets tied under dispersion. This is an algebraic example, not a benchmark measurement; it establishes a difference in scoring, not that the resulting ordering is better for classification. Conversely, dispersion removes the determinant-only zero when just two inputs coincide (Proposition 1). These two observations motivate testing the combined score against each component.

B.2 Score Properties, Normalization, and Scope

The results below explain the score used for evidence assessment in Section 4.3. Proposition 1 removes the determinant's ambiguous zero; Proposition 2 turns a small ARV into pairwise agreement; Proposition 3 states when that agreement bounds prototype perturbation. The final example separates these properties from the empirical behavior of the fusion search.

All vectors in this section have unit Euclidean norm. Write $G_{ij} = \exp(-\|v_i - v_j\|_2^2/(2\sigma^2))$, $V_{\text{det}} = \sqrt{\det G}$, and $D_k = \frac{1}{3}\sum_{i<j}(1 - G_{ij})$ for three inputs. The RBF Gram matrix is positive semidefinite, so its determinant is nonnegative in exact arithmetic; the implementation clips small negative numerical errors.

Proposition 1 (Zero-score characterization). For $\sigma > 0$ and $\lambda_{\text{disp}} > 0$,

$$S_{\text{geo}} = V_{\text{det}} + \lambda_{\text{disp}} D_k = 0 \quad \Longleftrightarrow \quad v_1 = v_2 = v_3. \tag{13}$$

Proof. Both summands are nonnegative. A zero score therefore implies $D_k = 0$. Every term $1 - G_{ij}$ is nonnegative, so each is zero. For an RBF kernel, $G_{ij} = 1$ if and only if $v_i = v_j$. Conversely, coincident inputs give an all-ones Gram matrix and zero dispersion. □

This property rules out a zero score caused solely by two coincident inputs. It does not imply that the score orders candidate classification accuracy correctly. In particular, the dispersion term alone has the same zero set.

Proposition 2 (Pairwise mismatch). If $S_{\text{geo}} \le \varepsilon < \lambda_{\text{disp}}/3$, every pair satisfies

$$\|v_i - v_j\|_2 \le r = \min\left\{2,\ \sigma\sqrt{-2\log(1 - 3\varepsilon/\lambda_{\text{disp}})}\right\}. \tag{14}$$

Proof. Since $D_k \le \varepsilon/\lambda_{\text{disp}}$, each of its three nonnegative summands is at most $3\varepsilon/\lambda_{\text{disp}}$. Inverting the RBF kernel gives the second bound; unit vectors are at most distance two apart. For larger ε, only that trivial bound is available from this argument. □

Proposition 3 (Normalized prototype perturbation). Let $X_1, \ldots, X_K$ be independent, identically distributed unit-norm support features of one class, $\mu = \mathbb{E}[X_i]$, $\|\mu\|_2 \ge \gamma > 0$, and

$u = \mu/\|\mu\|_2$ its population direction. Assume the empirical mean $\bar{X} = K^{-1}\sum_i X_i$ is nonzero almost surely and write $p^{\text{real}} = \text{norm}(\bar{X})$. For $0 < \delta < 1$, define

$$h_K(\delta) = \frac{1 + \sqrt{2\log(1/\delta)}}{\sqrt{K}}. \tag{15}$$

On an event of probability at least $1-\delta$ over the support sample, every generated prototype satisfying the source-triplet condition in Proposition 2 with $r < 2$, and every $\alpha \in [0, 1]$, obey

$$\left\|\text{norm}\big((1-\alpha)p^{\text{gen}} + \alpha p^{\text{real}}\big) - u\right\|_2 \le e_c := \min\left\{2,\ 2(1-\alpha)r + \frac{2h_K(\delta)}{\gamma}\right\}. \tag{16}$$

The population center here is a *direction*, not an unnormalized population mean. The condition on γ is necessary to control the effect of normalization.

Proof. Independence and bounded norms imply $\mathbb{E}\|\bar{X} - \mu\|_2 \le K^{-1/2}$. Replacing one support feature changes this norm by at most $2/K$. The bounded-differences inequality therefore yields $\|\bar{X} - \mu\|_2 \le h_K(\delta)$ with probability at least $1 - \delta$ (McDiarmid, 1989). For any nonzero a, b,

$$\|\,\text{norm}(a) - \text{norm}(b)\|_2 \le \frac{2\|a - b\|_2}{\|b\|_2}, \tag{17}$$

which follows by adding and subtracting $a/\|b\|_2$ and applying the reverse triangle inequality. Taking $a = \bar{X}, b = \mu$ gives $\|p^{\text{real}} - u\|_2 \le 2h_K(\delta)/\gamma$. Let $z = (1-\alpha)p^{\text{gen}} + \alpha p^{\text{real}}$. Proposition 2 gives $\|z - p^{\text{real}}\|_2 \le (1-\alpha)r$. The condition $r < 2$ excludes antipodal endpoints and ensures $z \neq 0$. Applying Eq. (17) with $b = p^{\text{real}}$ and then the triangle inequality proves Eq. (16). The support concentration event is shared by all candidate paths and interpolation weights, so selection based on that same support set does not require assuming independence of the selected generator and the supports. □

Classification consequence and its limits. For a unit query q, $\|p_c - u_c\|_2 \le e_c$ implies $|q^\top p_c - q^\top u_c| \le e_c$. Let $\beta_c = \lambda_{\text{conf}}\mathbb{I}[c \notin \mathcal{C}_0]\widetilde{S}_c$ be the implemented reliability penalty. An ideal winning class y remains the winner if, for every competing class j,

$$q^\top(u_y - u_j) > e_y + e_j + \beta_y - \beta_j. \tag{18}$$

This follows by lower-bounding the actual penalized logit difference. A simultaneous probability statement for multiple classes uses per-class failure probabilities whose sum is at most the desired total failure probability. The bound is necessarily vacuous for the five-support setting used in our experiments. For $K = 5$ and $\delta = 0.05$, $h_K(\delta) = (1 + \sqrt{2\log 20})/\sqrt{5} \approx 1.54$, while unit-norm features imply $\gamma \le \|\mu\|_2 \le 1$, so $2h_K(\delta)/\gamma \ge 3.08 > 2$ and $e_c = \min\{2, \ldots\} = 2$ coincides with the trivial diameter bound for unit vectors. The bound becomes informative only for substantially larger K or a population mean with norm bounded away from zero; it motivates agreement control but does not establish an accuracy gain, reliability calibration, or the necessity of the determinant term.

Endpoint behavior is a separate empirical question. At $\alpha = 1$, the fusion inputs are $(t, p^{\text{real}}, p^{\text{real}})$, so the determinant vanishes while

$$S_{\text{geo}} = \frac{2\lambda_{\text{disp}}}{3}\left(1 - k(t, p^{\text{real}})\right). \tag{19}$$

This is positive unless $t = p^{\text{real}}$, but can still be the minimum over the interpolation grid. For example, take $t, p^{\text{real}}, p^{\text{gen}}$ to be three orthogonal unit vectors, $\sigma = 0.5$, and $\lambda_{\text{disp}} = 1$. If the base penalty is inactive, the 21-point grid still selects $\alpha = 1$ (score approximately 0.65446). Thus the observed reduction of boundary selections in Figure 6 is an experimental observation, not a consequence of Proposition 1. This constructed example is a mathematical check and is not included in benchmark accuracy results.

C Component Controls

This appendix consolidates selection and fusion controls, compares ARV with simpler scores under matched candidate pools, and reports a joint-stage comparison. Session-wise and additional accuracy results appear in Appendix D.

C.1 Reported Selection and Fusion Controls

Table 4 brings together existing replacement measurements from Table 3 and Appendix D. For miniImageNet, the replacement rows use an earlier penalty-enabled setting; the full-method reference is the penalty-enabled row (72.91%) in Table 3, not its correction-off core row. Available CIFAR-100 component records do not preserve enough run metadata to establish the same correction setting or paired reference. The displayed gaps are descriptive contrasts, not matched estimates of component effects under the main setting; they are not additive.

Table 4: Reported component controls. Values consolidate existing manuscript results; the difference column is the reported full-method value minus the replacement value.

Benchmark	Replacement	Avg (%)	Full (%)	Difference
miniImageNet	Random selection ($M = 3$)	72.14	72.91	+0.77
CIFAR-100	Random selection ($M = 3$)	75.51	75.88	+0.37
miniImageNet	Fixed fusion ($\alpha = 0.5$)	71.11	72.91	+1.80
CIFAR-100	Fixed fusion ($\alpha = 0.5$)	74.68	75.88	+1.20
miniImageNet	No base-class regularization	72.13	72.91	+0.78
CIFAR-100	No base-separation penalty	75.30	75.88	+0.58
CIFAR-100	Determinant-only fusion	74.91	75.88	+0.97

Selecting which evidence to use. The table reports a 0.77-point difference on miniImageNet and a 0.37-point difference on CIFAR-100 between the listed full-method and random-selection measurements. Both select from three candidates, holding candidate count fixed. On miniImageNet, the measurements are not a matched estimate of the selection effect under the main setting; the available CIFAR-100 records do not establish a matched run either.

Choosing the contribution of that evidence. The listed fixed-fusion measurements are 1.80 points below the full reference on miniImageNet and 1.20 points below on CIFAR-100; the no-base-class-regularization measurements are 0.78 and 0.58 points below, respectively. They are descriptive cross-run contrasts; the available records do not establish matched correction settings for these comparisons.

Regularizing the fusion score. In the reported CIFAR-100 comparison, adding dispersion to determinant-only fusion yields a 0.97-point higher average accuracy. The boundary diagnostic in Figure 6 records the accompanying reduction in support-only selections across five support-seed replicates. The matched-score controls below examine the determinant's contribution relative to dispersion alone.

Scope of the evidence. The table reports aggregate outcomes without inferred uncertainty intervals. The matched-candidate comparisons below separately vary the selection score and fusion score while retaining the remaining pipeline choices.

C.2 Matched-Candidate Score Controls

Table 5 separates the two uses of the score. ARV entries give the full-method reference, and each remaining entry evaluates the corresponding selector or fusion rule on the same candidate pool. Random selection and fixed fusion from Table 4 answer different questions and are therefore reported separately.



Figure 6: Boundary-selection distributions for determinant-only fusion and complete ARV over five support-seed replicates. With the dispersion term, the observed frequency of the large $\alpha = 1$ mode decreases across all three benchmarks.

Table 5: Four-score comparison matrix for average accuracy (%). Selection columns vary only the selector; fusion columns vary only the fusion score. ARV entries use the full-method reference.

	Selection score		Fusion score	
Score	CIFAR-100	miniImageNet	CIFAR-100	miniImageNet
ARV	75.88	72.61	75.88	72.61
Determinant only	75.55	72.25	74.91	71.76
Dispersion only	75.96	72.53	75.92	72.50
Mean cosine disagreement	75.78	72.66	75.70	72.65

Alternative agreement scores. To test the score components, define lower-is-better alternatives on the same normalized triplet (t, r, g):

$$S_{\text{disp}} = D_k(\{t, r, g\}), \tag{20}$$

$$S_{\cos} = \tfrac{1}{3}[(1 - t^\top r) + (1 - t^\top g) + (1 - r^\top g)], \tag{21}$$

$$S_{tr} = 1 - t^\top r, \qquad S_{gr} = 1 - g^\top r. \tag{22}$$

S_{disp} measures dispersion alone, while $S_{\cos}$ compares ARV with mean cosine disagreement. The two-source alternatives S_{tr} and S_{gr} are defined but absent from the reported table. The available results therefore cannot establish whether all three sources improve on a two-source score. The zero-score property alone does not determine the empirical ordering.

Observed selection results. Relative to ARV selection, determinant-only selection is lower by 0.33 points on CIFAR-100 and 0.36 points on miniImageNet. Dispersion-only selection differs by $+0.08$ and -0.08 points, while mean cosine disagreement differs by -0.10 and $+0.05$ points. ARV and these simpler selectors are close, with dataset-dependent ordering.

Observed fusion results. With the ARV-selected candidate fixed, the reported determinant-only values are lower than the ARV reference by 0.97 points on CIFAR-100 and 0.85 points on miniImageNet. Dispersion-only fusion differs by $+0.04$ and -0.11 points, while cosine disagreement differs by -0.18 and $+0.04$ points. These are single-run point estimates: they disfavor the unregularized determinant in the reported comparison, while leaving the ordering of the complete score and simpler dispersion/cosine alternatives unresolved.

Paired comparison. Figure 4 reports eight paired replicates for fusion-score comparison. Each replicate contributes its mean accuracy over sessions; the large points average these values across the eight replicates, giving 75.47% for ARV. By contrast, the main benchmark result of 75.66% in Table 2 is the session average from a single run, not the eight-replicate mean. Paired differences and confidence intervals in the figure are computed solely from its paired replicates. The paired ARV-minus-determinant interval is positive, whereas the dispersion and cosine intervals include zero. This supports regularization of the determinant while leaving the ordering of the simpler alternatives unresolved in this comparison.

C.3 Joint-Stage Effects and Interaction

Let A_{ga} denote miniImageNet average accuracy with grounding $g \in \{0, 1\}$ and assessment-guided selection/fusion $a \in \{0, 1\}$. The factorial comparison in Table 6 distinguishes within-grid stage effects from their interaction.

Table 6: Joint-stage comparison on miniImageNet (average accuracy, %). All cells use three candidates; auxiliary base-separation and reliability penalties are disabled.

	Random selection, fixed fusion	ARV selection and fusion
Class-name conditions	68.44	70.01
GIP conditions	70.24	71.81

Each cell uses $M = 3$ descriptions. Within each grounding setting, the assessment settings share the generated candidate pool. Supports, encoder, diffusion model, noise-seed schedule, and queries are matched across cells. Fixed fusion uses $\alpha = 0.5$; base-separation and classification-reliability penalties are disabled throughout. The full method is reported separately because it includes these auxiliary terms.

The grounding effect with assessment enabled is $A_{11} - A_{01}$; the assessment effect with grounding enabled is $A_{11} - A_{10}$. The interaction is

$$I = (A_{11} - A_{10}) - (A_{01} - A_{00}). \tag{23}$$

A positive I indicates a larger assessment gain under grounded conditions. Here, grounding improves accuracy by 1.80 percentage points under either assessment setting, while assessment improves accuracy by 1.57 points under either grounding setting. Thus, $I = 0.00$ percentage points at the reported precision: both stages contribute positively within this grid, without an observed positive interaction. This descriptive contrast has no across-seed uncertainty estimate and does not establish statistical equivalence or an exactly zero population interaction. It is distinct from the one-component replacements above, whose auxiliary policies differ.

D Additional Accuracy and Ablation Results

Table 7 supplements the overall comparison in Section 5. The subsequent CIFAR-100 analyses follow the same two questions as the main experiments: how grounding changes the conditions and how assessment changes the use of generated evidence. Broader method rows provide benchmark context; the component measurements remain descriptive where run settings are not matched.

Table 7: miniImageNet session-wise top-1 accuracy (%) for representative FSCIL methods. The last column reports the final-session gain of exhaustive GIP-Geo over each method.

Method	Venue	Session accuracy (%)									Avg	Final gain
		S0	S1	S2	S3	S4	S5	S6	S7	S8		
iCaRL (Rebuffi et al., 2017)	CVPR 2017	61.31	46.32	42.94	37.63	30.49	24.00	20.89	18.80	17.20	33.29	+45.41
TOPIC (Tao et al., 2020)	CVPR 2020	84.85	71.10	66.23	63.10	59.63	54.59	52.84	50.54	49.18	61.34	+13.43
CEC (Zhang et al., 2021)	CVPR 2021	84.85	74.68	69.81	66.68	63.21	58.17	56.42	54.12	52.76	64.52	+9.85
FACT (Zhou et al., 2022)	CVPR 2022	84.85	76.55	71.68	68.55	65.08	60.04	58.29	55.99	54.63	66.18	+7.98
MetaFSCIL (Chi et al., 2022)	CVPR 2022	72.04	67.94	63.77	60.29	57.58	55.16	52.90	50.79	49.19	58.85	+13.42
Replay (Liu et al., 2022)	ECCV 2022	71.84	67.12	63.21	59.77	57.01	53.95	51.55	49.52	48.21	58.02	+14.40
CLOM (Zou et al., 2022)	NeurIPS 2022	84.85	78.42	73.55	70.42	66.95	61.91	60.16	57.86	56.50	67.85	+6.11
NC-FSCIL (Yang et al., 2023)	ICLR 2023	84.02	76.80	72.00	67.83	66.35	64.04	61.46	59.54	58.31	67.82	+4.30
SAVC (Song et al., 2023)	CVPR 2023	81.12	76.14	72.43	68.92	66.48	62.95	59.92	58.39	57.11	67.05	+5.50
OrCo (Ahmed et al., 2024)	CVPR 2024	83.30	75.32	71.52	68.16	65.62	63.11	60.20	58.82	58.08	67.12	+4.53
CLOSER (Oh et al., 2024)	ECCV 2024	76.02	71.61	67.99	64.69	61.70	58.94	56.23	54.52	53.33	62.78	+9.28
Tri-WE (Lee et al., 2025)	CVPR 2025	84.13	81.41	76.65	73.59	70.10	65.13	63.42	61.02	60.13	70.62	+2.48
CD-FSCIL (Kang et al., 2025)	arXiv 2025	84.85	81.33	76.46	73.33	69.86	64.82	63.07	60.77	59.41	70.43	+3.20
GIP-Geo (exhaustive)	Ours	84.85	82.93	78.36	75.43	72.26	67.42	65.87	63.77	62.61	**72.61**	–

Grounding the generation conditions. On CIFAR-100, the reported class-name-only, support-conditioned, and support-conditioned-with-neighbors measurements are 74.08%, 75.40%, and 75.88%, respectively. Their point-estimate differences are 1.32 and 0.48 points and follow the direction of the miniImageNet measurements in Section 5.2. The first contrast changes the condition-construction strategy, while the second examines base-class context within GIP. Both measure downstream classification, not textual factuality in isolation. These are single-run measurements; the available records do not establish across-seed uncertainty or correction-policy matching to the main table.

Retrieved neighbors are selected by class-name similarity. They may share taxonomy or context without being visually confusable with the support examples. The aggregate ablation does not show whether any neighbor harmed generation. Future tests could compare this retrieval strategy with visual filtering.

CIFAR-100 component comparisons. For candidate selection, one, two, and three candidates obtain 75.40%, 75.66%, and 75.88%, while random selection among three candidates reaches 75.51%. For prototype construction, replacing ARV-guided fusion with a fixed $\alpha = 0.5$ average gives 74.68%, and removing base-class separation gives 75.30%. A separate classification-penalty comparison records 75.88% with and 75.66% without the correction. These two values are not a with/without comparison of the main-table run, which uses no correction; they should be read only within the separate ablation record. The reported component values are descriptive single-run comparisons.

Sensitivity and optimization. In the reported single-run measurements, increasing the number of generated exemplars per candidate from $N = 1$ to $N = 5$ and $N = 10$ changes accuracy from 73.12% to 75.21% and 75.88%. RBF bandwidths $\sigma = 0.1$, 0.5, and 1.0 give 74.95%, 75.88%, and 75.62%, respectively. Direct gradient optimization reaches 76.05%, compared with 75.88% for the gradient-free sweep; this 0.17-point difference is descriptive and has no across-seed uncertainty estimate.

Non-degeneracy. In the reported CIFAR-100 single-run comparison, fusion accuracy is 74.91% with determinant-only scoring and 75.88% with complete ARV. Figure 6 expands the boundary analysis to five support-seed replicates per dataset. Determinant-only fusion selects $\alpha = 1$ for 46.5%, 44.0%, and 50.4% of class-seed pairs on CIFAR-100, miniImageNet, and CUB-200, respectively;

complete ARV reduces these rates to 1.5%, 1.5%, and 1.0%. The replicated diagnostic measures the boundary-selection rate, not the accuracy difference.

D.1 Paired Support-Seed Comparison

We evaluate the core GIP-Geo pipeline with reliability correction disabled in eight paired CD-FSCIL/GIP-Geo runs per dataset, using support-seed identifiers 101, 202, 303, 404, 505, 606, 707, and 808. For each run, A_{avg} first averages all session accuracies, including Session 0; A_T is final-session accuracy. We then summarize these run-level values across seeds. Each gain is computed within a matched seed before aggregation. Thus, sessions are not treated as independent replicates.

Table 8: Paired support-seed results with reliability correction disabled ($n = 8$ pairs per dataset). Accuracy columns report mean $\pm$ sample standard deviation (SD) in percent; gain reports the mean $\pm$ SD of paired GIP-Geo-minus-CD-FSCIL differences in percentage points. SD describes the observed across-seed spread, not a confidence interval.

Dataset	Metric	CD-FSCIL	GIP-Geo	Paired gain
CIFAR-100	A_{avg}	71.61 ± 0.26	74.23 ± 0.21	$+2.62 \pm 0.15$
	A_T	65.22 ± 0.45	69.95 ± 0.44	$+4.73 \pm 0.28$
miniImageNet	A_{avg}	69.11 ± 0.30	71.67 ± 0.22	$+2.55 \pm 0.28$
	A_T	62.76 ± 0.58	67.03 ± 0.36	$+4.28 \pm 0.54$
CUB-200	A_{avg}	78.19 ± 0.24	82.30 ± 0.05	$+4.11 \pm 0.23$
	A_T	71.27 ± 0.30	78.37 ± 0.16	$+7.10 \pm 0.29$

Without reliability correction, GIP-Geo has higher session-average and final-session accuracy in all eight pairs on each dataset. These results support the core pipeline across the evaluated support seeds, not the benefit of the reliability extension or robustness to alternative pretraining or class orders. This run set is reported separately from the single-run measurements in Table 2 and the fusion-score comparison in Figure 4; their values are not pooled.

E Class-Level Reliability Diagnostics

This appendix examines source ARV S_c^* as a class-level reliability signal (Section 4.4), through score–accuracy association, class ordering, and classification correction.

Does ARV track differences in class difficulty? Figure 7 reports Spearman correlations of -0.942 on CIFAR-100 ($n = 40$), -0.899 on miniImageNet ($n = 40$), and -0.861 on CUB-200 ($n = 100$), with nominal $p < 10^{-4}$. For each novel class, accuracy is computed over its test queries at the session when that class is introduced; reliability correction is disabled. Classes with larger measured inconsistency tend to have lower accuracy, motivating the reliability extension. Each point concerns a class; classes share a model and incremental evaluation, so the nominal p-values do not adjust for possible dependence among points. We treat these as descriptive associations: they do not establish correct ordering of candidate generations within a class or demonstrate that applying the classification penalty improves accuracy.

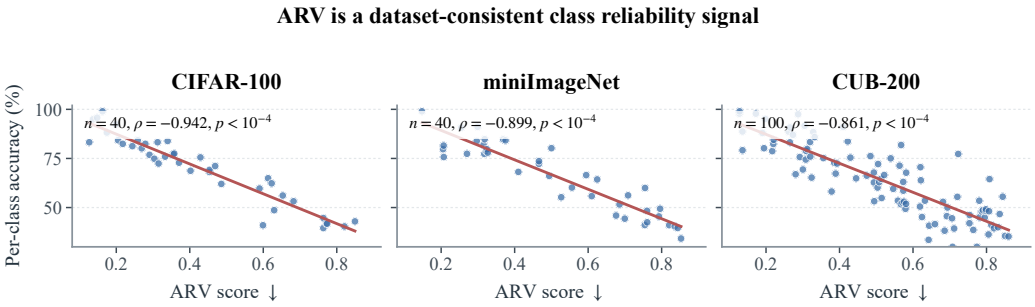

Figure 7: Class-level ARV–accuracy association on three benchmarks, measured without reliability correction. Each point is a novel class; accuracy is top-1 accuracy over that class's test queries at its introduction session. Each panel reports the number of classes, Spearman correlation, and nominal significance.

How should the score-ordering plots be interpreted? In Figure 8, class accuracy decreases from the lowest to the highest ARV quintile on all three datasets. Reported risk also rises as coverage expands from lower- to higher-score classes. These trends support using source ARV to distinguish class-level reliability; they do not define a query-wise uncertainty estimate or a rejection rule.

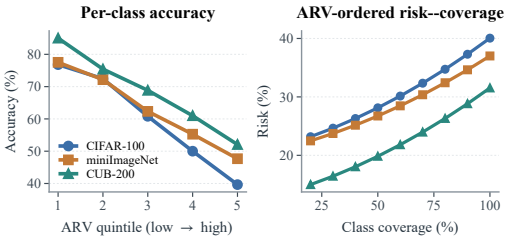

Figure 8: Left: mean per-class accuracy within ARV quintiles. Right: class-level risk (100% minus mean accuracy) among the lowest-ARV classes retained at each coverage. Accuracy uses each class's introduction-session queries; reliability correction is off. One run per dataset.

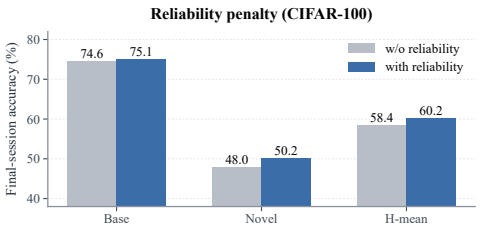

Figure 9: One separate CIFAR-100 run. Bars show final-session top-1 accuracy on 6,000 base and 4,000 novel queries, plus their harmonic mean, with and without the reliability penalty. Query counts are denominators, not run replicates.

Does reusing the assessment change classification performance? In the Figure 9 run, the penalty changes base, novel, and harmonic-mean accuracy from 74.6%/48.0%/58.4% to 75.1%/50.2%/60.2%. On support seed 101, novel accuracy falls from 65.92% to 18.33%. The two runs show different outcomes; they do not estimate how often the correction helps.

E.1 Class-Level Gains and Losses Across Runs

How can novel accuracy improve despite class suppression? The prediction export for Figure 9 contains 10,000 queries, with 100 per class. Eight novel classes exceed the exclusion threshold $2/4.89$; their correct predictions fall from 384 to zero. The other 32 novel classes gain 472 correct predictions, increasing from 1,536 to 2,008. Their net gain of 88 across 4,000 novel queries accounts for the 2.2-percentage-point improvement. Base-class correct predictions increase from 4,476 to 4,506 across 6,000 queries. Aggregate improvement therefore coexists with complete suppression of some classes.

Does the correction help every evaluated run? A separate CIFAR-100 diagnostic for support seed 101 contains 3,000 queries, with 30 per class. Applying the coefficient 4.89 raises base accuracy from 73.94% to 77.33%, while novel accuracy falls from 65.92% to 18.33% and overall accuracy from 70.73% to 53.73%. Seventeen of the 40 novel classes exceed the exclusion threshold $2/4.89$; their combined correct predictions fall from 341 to zero. Seed 101 therefore provides an observed instance of the suppression risk discussed above.

Table 9: Decomposition of the change in correct novel-class predictions. Excluded classes exceed $2/4.89$; changes are with penalty minus without penalty. Counts are within-run quantities with different query denominators, not directly comparable effect sizes.

Run	Novel queries	Excluded classes	Excluded change	Other change	Net change
Figure 9	4,000	8/40	-384	$+472$	$+88$
Seed 101	1,200	17/40	-341	-230	-571

The balance of gains and losses differs between runs (Table 9). In seed 101, some classes below the exclusion threshold also lose correct predictions, so the excluded-class count alone does not explain the outcome. The decomposition accounts for each run's accuracy change but cannot explain why score and margin distributions differ. The multi-seed comparison (Appendix D.1) disables correction; its results do not measure the penalty's across-seed benefit.

F Proxy and Resource Diagnostics

This appendix evaluates the efficiency extension in Section 4.4: how closely the proxy approximates exhaustive ARV selection and how much generation it avoids. Exhaustive selection remains the reference for both accuracy and decision agreement.

Ranking fidelity. On CIFAR-100, the full proxy reaches 75.87%, compared with 75.88% for exhaustive evaluation and 75.83% for Gaussian NLL training without the ranking term. Across 40 class-level candidate sets with $M = 3$, the audit yields 82.5% selection agreement, 97.5% top-2 recall, and mean within-class Spearman correlation $\rho = 0.863$. These measurements quantify recovery of the exhaustive ARV ranking. The paired proxy-minus-exhaustive accuracy difference is -0.01 ± 0.00 percentage points over Sessions 0–8; the spread is across sessions, not independent support seeds. The class-level audit and accuracy comparison are descriptive results from the reported run, and close aggregate accuracy with imperfect ranking agreement does not establish performance equivalence under repeated sampling.

Generation cost. The proxy matches the generated-image count of CD-FSCIL while retaining M=3 LLM calls per class. Table 10 separates analytical resource counts from the reported runtime ratio on CUB-200 (100 novel classes, N=10). The exhaustive and proxy variants both make M=3 LLM calls per class, but exhaustive evaluation generates images for all three descriptions, whereas the proxy generates images only for the selected description. The reported proxy runtime is approximately 1.3× the CD-FSCIL reference, compared with 3.0× for exhaustive evaluation. Available records omit the hardware, absolute elapsed times, and timer boundaries; these run-specific ratios do not support reproducible latency estimates. They may also depend on provider-specific LLM latency.

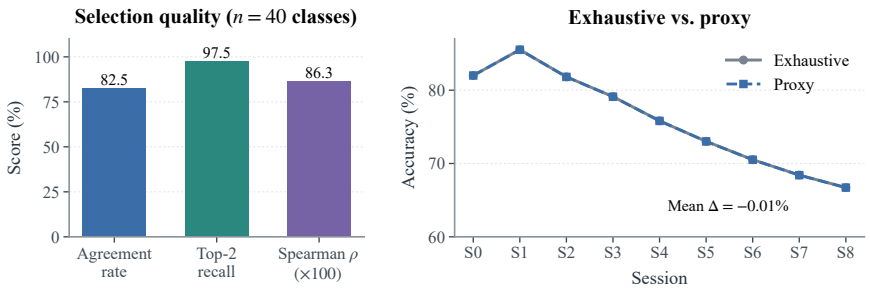

Figure 10: Score-proxy ranking audit over 40 class-level candidate sets. Left: 82.5% exact agreement, 97.5% top-2 recall, and 0.863 mean within-class Spearman correlation. Right: paired session-wise exhaustive-versus-proxy accuracy.

Table 10: Resource accounting on CUB-200 (100 novel classes, N=10). LLM calls and generated-image counts are analytical; relative time and peak memory are reported alongside them.

Method	Updated module	LLM calls	Images	Time	Memory (GB)
NC-FSCIL (Yang et al., 2023)	Classifier adaptation	–	–	2.5×	8.1
CD-FSCIL (Kang et al., 2025)	None	100	1,000	1.0×	3.6
GIP-Geo (M=3 exhaustive)	None	300	3,000	3.0×	3.6
GIP-Geo (proxy)	Proxy only	300	1,000	1.3×	3.6