MEASUREMENT-DRIVEN ISOMETRIC SHARED-PROMPT TRANSPORT FOR GENERAL CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

General continual learning (GCL) requires single-pass learning from non-stationary streams without explicit task boundaries. Prompt tuning enables parameter-efficient adaptation of pretrained models, but recurring classes and changing class mixtures demand both knowledge reuse and distribution-specific adjustment. Selecting or combining independently learned prompts does not explicitly separate shared adaptation structure from distribution-specific variation, potentially dispersing reusable knowledge across prompt states under limited supervision. Meanwhile, imbalanced class observations can bias classifiers toward recent data, further limiting the effective use of historical knowledge. We propose MIST (Measurement-Driven Isometric Shared-Prompt Transport), a replay-free framework that coordinates structured prompt adaptation and adaptive classification through continuous historical associations. MIST uses frozen-feature statistics to associate inputs with historical distributions and condition isometric transformations of shared latent prompts, preserving token geometry across conditions for a given shared latent representation. The same associations weight temporal classifier ensembles, while statistical novelty and confidence advantage regulate their fusion with an online plastic classifier to balance knowledge retention and adaptation. Experiments demonstrate absolute gains of up to 12.10%, 9.14%, and 24.25% in average anytime accuracy over the strongest evaluated baselines on CIFAR-100, ImageNet-R, and CUB-200, respectively. Our source code is available at MIST.

1 INTRODUCTION

Continual learning (CL) (Wang et al., 2024; Dohare et al., 2024; Wang et al., 2022d) seeks to accumulate knowledge from evolving observations while retaining previously acquired capabilities. This ability is important for deployed intelligent systems, including autonomous agents and personalized perception systems, whose operating conditions change over time. General continual learning (GCL) considers single-pass, non-stationary streams without explicit task boundaries (Aljundi et al., 2019b;a; De Lange et al., 2021; Moon et al., 2023; Yan et al., 2026). Classes may recur at irregular intervals, and their relative frequencies can change as new categories appear (Fig.1.(a) and 2.(a)). Consequently, recent observations provide only a partial view of the recognition problem. GCL therefore requires timely adaptation, sustained knowledge retention, and efficient use of limited supervision.

Pretrained models (PTMs) provide a useful foundation for CL (Zhou et al., 2024; Wang et al., 2025). Prompt tuning adapts their representations through a small set of trainable parameters while keeping the backbone frozen, restricting optimization to lightweight adaptation modules (Jia et al., 2022; Wang et al., 2022d; Zhou et al., 2024). Existing approaches include input-dependent prompt selection (Wang et al., 2022d;b), complementary task-shared and task-specific prompts (Wang et al., 2022c), and weighted composition of prompt components (Smith et al., 2023) (Fig. 1(a)). Shared-prompt adaptation has also been explored under task imbalance (Hong et al., 2024). Recent GCL research further improves prompt initialization (Kang et al., 2025), incorporates imbalance-aware learning and output masking (Moon et al., 2023; Kang et al., 2025), and introduces temporal classifier ensembles (Yan et al., 2026). These advances support replay-free continual learning; never-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

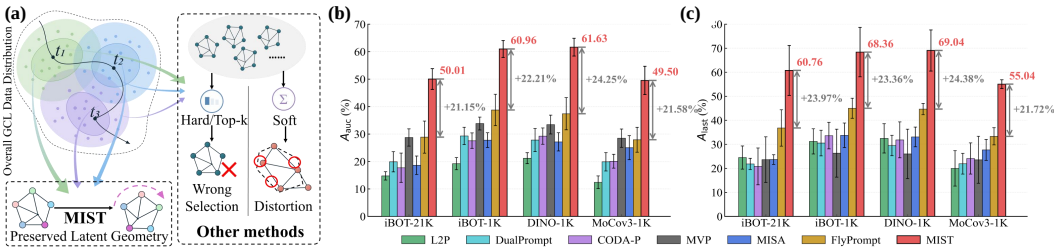

Figure 1: Motivation and performance of MIST. (a) Potential selection mismatch and geometric variation under prompt composition motivate conditional adaptation of a shared latent prompt. MIST preserves latent token geometry across conditions at a fixed shared configuration. (b,c) Average anytime and final accuracy on CUB-200 with self-supervised backbones. Error bars denote standard deviations over five seeds; annotated gains are absolute differences from FlyPrompt.

theless, freezing the backbone alone does not eliminate interference and forgetting in the remaining trainable prompts and classifiers (Kang et al., 2025; Qiao et al., 2025).

Addressing these limitations calls for examining how accumulated knowledge is reused in both prompt adaptation and classification as the data stream evolves. A central difficulty is reusing adaptation knowledge across recurring classes and evolving class mixtures (Moon et al., 2023). Prompt selection and composition distribute adaptation across multiple prompt states (Wang et al., 2022d; Smith et al., 2023), potentially dispersing related supervision across separate parameter sets. A shared prompt provides an alternative for accumulating common knowledge (Hong et al., 2024), but subsequent updates can disturb earlier adaptations (Qiao et al., 2025). The challenge is therefore to accommodate input-dependent variation within a reusable prompt structure: selection or linear composition alone does not guarantee preservation of token norms and pairwise relations across conditions (Fig. 1(a)). At the prediction level, uneven class exposure can bias online classifiers toward recently frequent classes (Moon et al., 2023; Kang et al., 2025). Output masking and temporal ensembles alleviate forgetting (Kang et al., 2025; Yan et al., 2026), but retaining historical states alone does not establish how their relevance should be weighed against current predictive evidence. These considerations motivate adapting a shared latent prompt structure across the entire stream through geometry-preserving conditional variation, while coordinating historical and plastic predictions according to historical relevance and current evidence.

To this end, we propose **MIST** (**M**easurement-Driven **I**sometric **S**hared-Prompt **T**ransport), a replay-free GCL framework that adapts a shared prompt structure across the entire stream. MIST separates a shared, learnable latent configuration from input-dependent variation through conditional orthogonal transformations. Continuous historical associations, derived from statistics accumulated in a frozen feature space, combine compact transformation coordinates without requiring task identities. For a fixed shared configuration, these transformations preserve latent token norms and pairwise inner products across input conditions. The same associations aggregate historical ensemble predictions, while statistical novelty and the confidence advantage of an online plastic classifier govern their fusion with plastic predictions. This common statistical reference coordinates structured prompt adaptation with the input-dependent use of historical and current evidence.

In summary,MIST provides a structured formulation of shared adaptation and historical prediction for replay-free GCL. Our main contributions are summarized as follows:

1. We propose conditional isometric transport of shared latent prompts, enabling adaptation knowledge to be reused across time through a common representation while accommodating distributional changes through input-dependent isometric transformations.

2. We develop an adaptive classification mechanism that uses the same historical associations to aggregate temporal classifier predictions and integrates them with an online plastic classifier according to statistical novelty and confidence advantage.

3. We systematically evaluate MIST across supervised and self-supervised pretrained backbones, demonstrating improvements in online recognition and knowledge retention under recurring classes and changing class mixtures, with particularly pronounced gains on self-supervised backbones (Fig.1.(b,c)).

2 RELATED WORK

Continual learning. Continual learning aims to acquire knowledge from sequentially arriving data while retaining previously learned capabilities (Wang et al., 2024; Dohare et al., 2024; Wang et al., 2022d). Existing approaches are commonly categorized into replay, regularization, and parameter isolation (Wang et al., 2024). Replay-based methods retain or generate historical samples for joint training with incoming data (Ostapenko et al., 2019; Buzzega et al., 2020; Wang et al., 2022a; Kang et al., 2023). Regularization-based methods constrain updates to important parameters or preserve the outputs of earlier models to reduce interference with acquired knowledge (Wu et al., 2019; Cha et al., 2021; Douillard et al., 2020). Parameter-isolation methods allocate dedicated parameters, subnetworks, or masks to restrict interference between task updates (De Lange et al., 2021; Serra et al., 2018). General continual learning (GCL) further considers single-pass, non-stationary streams without explicit task boundaries, where classes may recur and their composition and frequencies change over time (Mi et al., 2020; De Lange et al., 2021; Moon et al., 2023; Kang et al., 2025; Yan et al., 2026). A representative setting, Si-Blurry (Moon et al., 2023), combines classes restricted to individual stages with classes that can recur across stages. Randomized sample allocation produces changing class mixtures and imbalanced distributions, providing a concrete protocol for evaluating continual adaptation under these conditions (Fig.1.(a) and 2.(a)).

Prompt-based general continual learning. Prompt learning introduces lightweight trainable parameters into frozen pretrained backbones for downstream adaptation (Lester et al., 2021; Li & Liang, 2021; Wang et al., 2022d; 2026). Existing approaches organize adaptation knowledge through shared prompts, condition-specific prompts, or their combination, and construct prompts through input-dependent selection or weighted composition (Smith et al., 2023). Research addressing GCL has developed along two complementary directions: improving initialization and online training objectives to support efficient prompt adaptation under limited supervision, and using output masking (Moon et al., 2023; Kang et al., 2025; Park et al., 2026) and classifier ensembles across multiple timescales to mitigate prediction bias and forgetting under class recurrence and uneven exposure. Nevertheless, evaluations across pretraining configurations report substantially lower accuracy for several prompt-based methods with self-supervised backbones than with the evaluated supervised initializations (Wang et al., 2023; 2025), particularly in fine-grained recognition (Yan et al., 2026). These findings indicate that the effectiveness of existing prompt-based GCL methods remains sensitive to pretraining conditions (Fig.1.b and c).

In contrast to approaches that select or combine multiple prompt states, MIST separates input-dependent variation from a shared latent configuration through geometry-preserving transformations, supporting knowledge reuse while coordinating historical predictions with current evidence.

3 METHOD

Following prior GCL studies (Moon et al., 2023; Kang et al., 2025; Yan et al., 2026), we adopt the Si-Blurry protocol (Moon et al., 2023). A detailed description of the data protocol is provided in Appendix A. Built on a frozen pretrained ViT, MIST comprises three components, as shown in Fig. 2: (i) statistical measurement in a fixed feature space (Sec.3.1), which establishes continuous associations between current inputs and accumulated historical statistics; (ii) isometric shared-prompt transport (Sec.3.2), which enables input-dependent adaptation through conditional isometric transformations of shared latent prompts; and (iii) adaptive prediction fusion (Sec.3.3), which integrates historical EMA classifier responses with online plastic predictions using the historical associations and current predictive evidence. Internal intervals are instantiated according to prescribed cumulative sample-count boundaries, independently of benchmark task identities; the schedule is detailed in Appendix B.

3.1 CONTINUOUS STATISTICAL MEASUREMENT

Class recurrence and evolving class mixtures in GCL create overlapping statistical structures across historical intervals, motivating continuous associations for shared prompt adaptation and historical classifier aggregation. However, statistics accumulated in a prompt-adapted feature space can become inconsistent with current representations as prompt parameters evolve(Li et al., 2024; Goswami et al., 2023). Without replay, historical observations cannot be re-encoded to correct

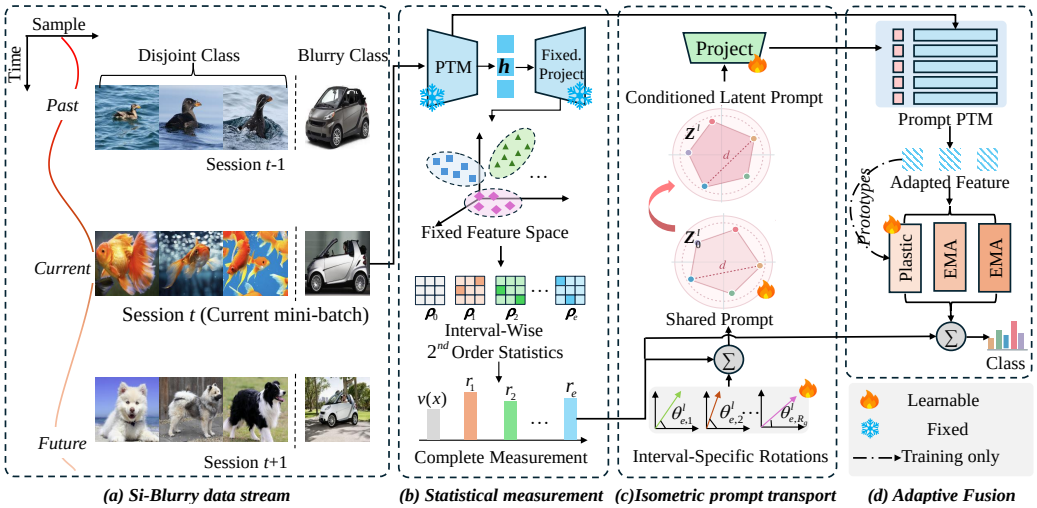

Figure 2: Overview of MIST. (a) A Si-Blurry stream with disjoint and recurring classes. (b) Fixed-feature statistics establish continuous historical associations. (c) These associations combine interval-specific transformation coordinates to condition a shared latent prompt. (d) EMA predictions are aggregated by a class-wise maximum within each interval and by association weights across intervals. Statistical novelty and confidence advantage regulate fusion with the plastic classifier. Detached class prototypes supervise the plastic classifier during training; the active EMA bank is updated by temporal averaging.

this mismatch (Yu et al., 2020; Gomez-Villa et al., 2024). MIST therefore maintains interval-level second-order statistics in the frozen, unprompted feature space and evaluates each input against these accumulated structures. The resulting associations provide a consistent statistical reference for conditioning shared prompt transformations and aggregating historical predictions.

For an input $\mathbf{x}$, the frozen backbone extracts the unprompted class-token feature $\mathbf{h}(\mathbf{x}) = f_\Theta^{\text{cls}}(\mathbf{x}) \in \mathbb{R}^{1\times D}$. A fixed projection $\mathbf{R} \in \mathbb{R}^{D\times d_m}$, with $\mathbf{R}^\top\mathbf{R} = \mathbf{I}_{d_m}$, produces the normalized query $\mathbf{q}(\mathbf{x}) = \mathbf{h}(\mathbf{x})\mathbf{R}/\|\mathbf{h}(\mathbf{x})\mathbf{R}\|_2$. Keeping both feature extraction and projection fixed provides a consistent coordinate system for historical association statistics.

The E intervals instantiated so far are indexed by $e \in \{1,\dots,E\}$. For each interval e, $\mathcal{I}_e$ contains the indices of its observed samples, with $N_e = |\mathcal{I}_e|$. Feature outer products are accumulated online to construct a normalized second-order statistic:

$$\mathbf{C}_e = \frac{1}{\max(N_e,1)} \sum_{i\in\mathcal{I}_e} \mathbf{q}(\mathbf{x}_i)^\top \mathbf{q}(\mathbf{x}_i), \qquad \boldsymbol{\rho}_e = \frac{\mathbf{C}_e + \epsilon\mathbf{I}_{d_m}}{\mathrm{Tr}(\mathbf{C}_e + \epsilon\mathbf{I}_{d_m})}, \tag{1}$$

where $\epsilon > 0$ is a stabilizing constant. This representation retains second-order relations among feature directions within each interval using only an accumulated matrix and a sample count, without storing past examples. An interval without observations naturally yields an isotropic statistic.

To assess potentially overlapping interval statistics against a common reference, we introduce an isotropic component $\boldsymbol{\rho}_0 = \mathbf{I}_{d_m}/d_m$ with a fixed prior $\pi_0 \in (0,1)$. The remaining components receive smoothed count-based priors $\pi_e = (1-\pi_0)(N_e+1)/\sum_{j=1}^{E}(N_j+1)$. Following the standard pretty-good measurement (PGM) construction (Hausladen & Wootters, 1994; Childs & Van Dam, 2010; Eldar & Forney, 2002), the mixture statistic and measurement operators are

$$\mathbf{S} = \sum_{e=0}^{E} \pi_e\boldsymbol{\rho}_e, \qquad \mathbf{M}_e = \mathbf{S}^{-1/2}\pi_e\boldsymbol{\rho}_e\mathbf{S}^{-1/2}, \quad e = 0,\dots,E. \tag{2}$$

The common inverse-square-root transformation evaluates each component relative to the same mixture of second-order structures. The isotropic component ensures that $\mathbf{S}$ is positive definite, yielding $\mathbf{M}_e \succeq 0$ and $\sum_{e=0}^{E}\mathbf{M}_e = \mathbf{I}_{d_m}$. The measurement properties are proved in Appendix D.1.

A normalized query produces nonnegative responses $a_e(\mathbf{x}) = \mathbf{q}(\mathbf{x})\mathbf{M}_e\mathbf{q}(\mathbf{x})^\top$, satisfying $\sum_{e=0}^{E} a_e(\mathbf{x}) = 1$. The interval weights $r_e(\mathbf{x}) = a_e(\mathbf{x})/\sum_{j=1}^{E} a_j(\mathbf{x})$ express relative associa-

tions with accumulated interval statistics without discrete task assignment. The same association vector $\mathbf{r}(\mathbf{x})$ conditions shared prompt transformations and weights historical classifier aggregation. The isotropic response $\nu(\mathbf{x}) = a_0(\mathbf{x})$ places greater weight on directions weakly supported by the mixture statistic, providing a relative statistical novelty signal for fusing historical and online plastic predictions.

3.2 Measurement-Driven Isometric Shared-Prompt Transport

Conditional Isometric Transport. Class recurrence creates opportunities to reuse adaptation knowledge across intervals, while evolving class mixtures require input-dependent variation. Independent prompt sets can disperse reusable adjustments across separately optimized parameters, whereas an input-independent shared prompt does not explicitly represent such variation. Prompt selection and weighted composition enable conditional adaptation but do not, by themselves, preserve a common token geometry across conditions. Orthogonal transformations preserve pairwise inner products (Kornblith et al., 2019; Williams et al., 2021), whereas averaging rotations does not generally yield a rotation (Moakher, 2002). Motivated by geometry-preserving adaptation (Qiu et al., 2023), MIST separates a shared latent prompt configuration from compact interval-specific transformation coordinates, allowing conditional changes in orientation while retaining the configuration's internal geometry.

For each prompted Transformer layer $l \in \mathcal{L}$, we learn a shared latent prompt $\mathbf{Z}_0^l \in \mathbb{R}^{P\times d_p}$, comprising P tokens of dimension d_p. Interval e maintains coordinates $\boldsymbol{\theta}_e^l \in \mathbb{R}^{R_g}$, specifying rotation angles on R_g fixed coordinate planes (i_r, j_r), with $1 \le i_r < j_r \le d_p$. The Givens matrix $\mathbf{G}_r(\theta)$ performs a standard rotation on the corresponding plane and leaves all remaining coordinates unchanged. Conditioning weights $\mathbf{w}(\mathbf{x})$ determine the effective angles $\bar{\theta}_r^l(\mathbf{x}) = \sum_{e=1}^{E} w_e(\mathbf{x})\theta_{e,r}^l$, yielding

$$\mathbf{U}^l(\mathbf{x}) = \prod_{r=1}^{R_g} \mathbf{G}_r\big(\bar{\theta}_r^l(\mathbf{x})\big), \qquad \mathbf{Z}^l(\mathbf{x}) = \mathbf{Z}_0^l\mathbf{U}^l(\mathbf{x}). \tag{3}$$

The rotations follow a fixed order. At inference, $\mathbf{w}(\mathbf{x}) = \mathbf{r}(\mathbf{x})$; the training weights are specified below.

Because $\mathbf{U}^l(\mathbf{x})$ is orthogonal, every fixed shared configuration satisfies

$$\mathbf{Z}^l(\mathbf{x})\mathbf{Z}^l(\mathbf{x})^\top = \mathbf{Z}_0^l(\mathbf{Z}_0^l)^\top. \tag{4}$$

Thus, conditioning preserves latent token norms, pairwise inner products, and distances at a fixed shared configuration; the proof and sensitivity bound are provided in Appendix D.2. A shared learnable projection $\mathbf{W}^l \in \mathbb{R}^{d_p\times D}$ produces the backbone prompts through $\mathbf{P}^l(\mathbf{x}) = \mathbf{Z}^l(\mathbf{x})\mathbf{W}^l + \mathbf{B}$, where $\mathbf{B} \in \mathbb{R}^{P\times D}$ contains fixed positional embeddings. The invariant applies before this projection and across conditions at a fixed shared configuration; both $\mathbf{Z}_0^l$ and $\mathbf{W}^l$ remain trainable. Prompt tokens participate in self-attention within each selected block and are removed from its output. The resulting image tokens ultimately produce the adapted class-token feature $\mathbf{z}(\mathbf{x}) \in \mathbb{R}^D$.

Geometry Consolidation and Online Adaptation. Orthogonal transport preserves token relationships while rotating the second-order structure in latent coordinates. This orientation affects how the transported prompt interacts with the shared embedding map. We therefore regularize the active rotation coordinates against the structures induced by earlier interval coordinates, encouraging locally compatible transformations while retaining conditional variation.

With e_{cur} denoting the active internal interval, $\mathbf{Z}_j^l$ for $j < e_{\mathrm{cur}}$ denotes the current shared prompt transformed using the historical coordinates $\boldsymbol{\theta}_j^l$. Its reference second moment is $\mathbf{\Gamma}_j^l = \mathrm{sg}\big((\mathbf{Z}_j^l)^\top\mathbf{Z}_j^l / \max(\|\mathbf{Z}_j^l\|_F^2, \epsilon_p)\big)$, where $\epsilon_p > 0$ stabilizes normalization and sg stops gradients. These references are recomputed from the current shared prompt and historical transformation coordinates. The active coordinates define $\mathbf{A}_{e_{\mathrm{cur}}}^l = \sum_{r=1}^{R_g} \theta_{e_{\mathrm{cur}},r}^l \mathbf{B}_r$, where $\mathbf{B}_r$ has entries $(\mathbf{B}_r)_{i_r j_r} = 1$, $(\mathbf{B}_r)_{j_r i_r} = -1$, and zeros elsewhere. We penalize

$$\mathcal{L}_{\mathrm{con}} = \frac{\sum_{l\in\mathcal{L}}\sum_{j<e_{\mathrm{cur}}} \big\|\mathbf{A}_{e_{\mathrm{cur}}}^l\mathbf{\Gamma}_j^l - \mathbf{\Gamma}_j^l\mathbf{A}_{e_{\mathrm{cur}}}^l\big\|_F^2}{|\mathcal{L}|(e_{\mathrm{cur}}-1)d_p^2}, \tag{5}$$

with $\mathcal{L}_{\text{con}} = 0$ for the first interval. The commutator characterizes the first-order change of a reference moment under an infinitesimal rotation, providing a local compatibility constraint. Its local justification is provided in Appendix D.3. The quadratic penalty $\mathcal{L}_{\text{geo}} = (|\mathcal{L}|R_g)^{-1} \sum_{l\in\mathcal{L}} \|\boldsymbol{\theta}^l_{e_{\text{cur}}}\|_2^2$ additionally favors smaller angles. In the training objective, the contribution of $\mathcal{L}_{\text{con}}$ is modulated by the minibatch average of $1-\nu(\mathbf{x})$, assigning greater regularization weight to statistically familiar inputs.

A newly instantiated interval initially has limited statistical support, so its association weight may be small. Training therefore uses $\mathbf{w}^{\text{tr}}(\mathbf{x}) = (1-\beta)\mathbf{e}_{\text{cur}} + \beta\mathbf{r}(\mathbf{x})$, where $\mathbf{e}_{\text{cur}}$ is the current interval's one-hot indicator and $\beta \in [0,1]$ controls historical mixing. This explicitly weights the active interval while retaining contributions from accumulated statistics. At inference, conditioning uses only $\mathbf{r}(\mathbf{x})$, without an externally supplied interval identity.

3.3 Adaptive Prediction Fusion and Online Optimization

Historical Association and Adaptive Prediction Fusion. Structured prompt adaptation does not by itself prevent the classifier from being biased toward recent observations. In GCL, changing class coverage and imbalanced mini-batches can make immediate adaptation inconsistent with historical decisions. MIST therefore combines a plastic classifier with interval-specific temporal ensembles. Historical associations determine the contribution of each ensemble, while statistical novelty and the plastic branch's confidence advantage regulate the contribution of current evidence.

For the adapted feature $\mathbf{z}(\mathbf{x})$, the plastic classifier produces cosine logits $\ell_c^{\text{pl}}(\mathbf{x}) = s\,\mathbf{z}(\mathbf{x})^\top \mathbf{w}_c / (\|\mathbf{z}(\mathbf{x})\|_2 \|\mathbf{w}_c\|_2)$, where $\mathbf{w}_c$ is the weight vector of class c and $s > 0$ is a learnable scale. Its prediction is $\mathbf{p}_{\text{pl}}(\mathbf{x}) = \text{SoftMax}(\boldsymbol{\ell}^{\text{pl}}(\mathbf{x}))$. Each interval e additionally maintains M EMA cosine classifiers with distinct decay rates, evaluated on the same adapted feature. Their predictions $\mathbf{p}_{e,m}(\mathbf{x})$ are aggregated through the class-wise maximum $s_{e,c}(\mathbf{x}) = \max_m p_{e,m,c}(\mathbf{x})$. This operation retains responses across complementary temporal scales; the resulting vector $\mathbf{s}_e$ need not sum to one.

The historical associations weight the interval responses, and a sample-dependent coefficient $g(\mathbf{x}) \in [0,1]$ combines the historical and plastic branches:

$$\begin{aligned} \mathbf{s}_{\text{st}}(\mathbf{x}) &= \sum_{e=1}^{E} r_e(\mathbf{x})\mathbf{s}_e(\mathbf{x}), \\ \mathbf{s}_{\text{final}}(\mathbf{x}) &= g(\mathbf{x})\mathbf{p}_{\text{pl}}(\mathbf{x}) + \big[1-g(\mathbf{x})\big]\mathbf{s}_{\text{st}}(\mathbf{x}). \end{aligned} \tag{6}$$

The gate increases with the statistical novelty $\nu(\mathbf{x})$ and the positive confidence advantage $\Delta(\mathbf{x}) = [\max_c p_{\text{pl},c}(\mathbf{x}) - \max_c s_{\text{st},c}(\mathbf{x})]_+$. Thus, familiar inputs can retain a substantial historical contribution, while statistical novelty or stronger plastic responses increase the weight of current adaptation. The same historical associations connect prompt conditioning with classifier aggregation. Prediction selects the largest final response among observed classes, without requiring an interval identity. The exact gate and EMA update rules are provided in the Appendix E.1.

Online Optimization. We supervise both plastic and fused predictions so that online learning directly optimizes the combined decision rule. Class prototypes provide additional constraints for categories absent from the current mini-batch. For a mini-batch $\mathcal{B}$, the mask $\mathbf{m}$ assigns zero to retained classes and $-\infty$ otherwise. It retains current-batch classes, or all observed classes when batch masking is disabled. The sample-based objective is

$$\mathcal{L}_{\text{data}} = 0.7\,\mathbb{E}_{(\mathbf{x},y)\in\mathcal{B}}\,\text{CE}\big(\boldsymbol{\ell}^{\text{pl}}(\mathbf{x}) + \mathbf{m}, y\big) + 0.3\,\mathbb{E}_{(\mathbf{x},y)\in\mathcal{B}}\,\text{CE}(\log \mathbf{s}_{\text{final}}(\mathbf{x}) + \mathbf{m}, y)\,. \tag{7}$$

Here, expectations denote mini-batch averages. Applying cross-entropy to the logarithmic fused scores provides supervision without assuming that the temporal aggregation is a normalized probability distribution.

For each observed class, a prototype $\boldsymbol{\mu}_c$ accumulates the mean of normalized adapted features without storing past images. The prototype loss $\mathcal{L}_{\text{proto}}$ averages the plastic classifier's cross-entropy over all available detached prototypes, without class masking. Classification uses $\mathcal{L}_{\text{cls}} = \frac{1}{2}(\mathcal{L}_{\text{data}} + \mathcal{L}_{\text{proto}})$ when prototypes are available, and $\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{data}}$ otherwise. Together with the geometric regularizers, the online objective is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{con}}\bar{f}_{\mathcal{B}}\mathcal{L}_{\text{con}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}}, \tag{8}$$

Table 1: Overall performance of representative methods on three GCL benchmarks across different PTMs. Higher A_{auc} and A_{last} are better, while lower F_{last} is better.

PTM	Method	CIFAR-100			ImageNet-R			CUB-200		
		A_{auc} (%) ↑	A_{last} (%) ↑	F_{last} (%) ↓	A_{auc} (%) ↑	A_{last} (%) ↑	F_{last} (%) ↓	A_{auc} (%) ↑	A_{last} (%) ↑	F_{last} (%) ↓
Sup-21K	Seq FT	19.71±3.39	10.42±4.92	–	7.51±3.94	2.29±0.85	–	3.47±0.41	1.49±0.42	–
	Linear Probe	49.69±6.09	23.07±7.33	–	29.24±1.26	16.87±3.14	–	28.96±2.46	17.33±3.08	–
	Seq FT w/ SL	64.90±7.18	62.06±1.89	–	47.20±1.47	39.60±2.43	–	56.16±4.32	56.50±3.08	–
	L2P	76.23±2.73	79.11±1.43	11.53±1.44	44.40±1.03	42.03±1.72	19.16±1.92	64.30±2.18	61.42±2.13	29.15±3.19
	DualPrompt	76.04±3.32	76.62±0.74	11.42±0.91	46.13±1.94	40.80±1.04	18.24±4.34	65.03±2.24	62.43±1.78	27.18±5.91
	CODA-P	79.13±3.06	80.91±0.70	9.95±1.30	51.87±2.81	48.09±2.75	18.49±2.44	66.01±2.20	62.90±2.46	27.43±3.42
	MVP	67.74±4.96	63.22±0.69	33.19±2.33	39.50±1.41	32.63±3.95	43.39±4.93	54.69±3.14	50.07±3.86	47.61±5.73
	MISA	80.35±2.39	80.75±1.24	9.67±1.39	51.52±2.09	45.08±1.43	21.46±4.25	65.40±3.01	60.20±1.82	30.06±3.70
	FlyPrompt	82.65±2.56	85.96±0.70	5.03±1.16	**56.58**±1.47	**55.27**±0.91	13.42±1.74	70.64±2.85	73.40±1.88	10.97±1.52
	MIST	**82.98**±1.59	**86.61**±0.35	**4.03**±0.79	51.96±1.48	50.68±1.00	**7.59**±1.19	**76.98**±2.02	**80.58**±1.54	**5.32**±2.00
Sup-21K/1K	L2P	63.88±7.79	68.96±7.63	14.88±6.27	47.10±1.01	42.22±1.94	35.88±4.51	42.96±4.13	45.00±3.83	35.05±9.31
	DualPrompt	68.02±2.08	67.04±5.84	19.18±5.13	52.80±4.91	47.39±1.60	30.73±7.45	46.80±2.89	46.39±2.76	35.40±8.16
	CODA-P	69.29±2.52	69.47±7.19	19.82±6.85	51.20±1.76	44.30±1.50	35.58±5.52	44.66±2.73	45.18±4.50	35.73±8.95
	MVP	64.69±3.77	51.29±7.56	46.98±8.15	48.99±2.01	38.12±5.20	51.11±8.56	44.10±2.81	33.97±9.62	62.93±8.05
	MISA	62.91±7.96	67.99±7.41	11.38±5.76	50.87±1.69	47.75±5.87	28.66±8.19	42.76±2.33	44.05±1.94	33.95±8.63
	FlyPrompt	78.48±1.31	80.39±3.54	9.50±2.33	62.01±2.32	56.55±3.94	21.59±4.28	54.42±4.67	55.50±3.55	21.71±3.56
	MIST	**84.27**±1.63	**88.49**±2.25	**3.58**±0.79	**71.15**±1.31	**71.05**±0.82	**6.40**±1.43	**75.08**±2.53	**78.95**±6.08	**5.96**±6.06
iBOT-21K	L2P	56.82±8.42	67.61±8.76	12.58±6.24	35.97±1.62	36.95±2.44	34.17±7.54	14.76±1.53	24.51±4.82	19.09±8.05
	DualPrompt	66.06±4.55	67.14±8.60	19.22±4.99	42.48±1.62	35.91±0.88	37.96±9.50	19.90±3.68	21.84±2.35	30.18±8.79
	CODA-P	62.13±7.17	63.38±7.98	22.20±5.61	45.50±1.66	39.44±1.35	41.21±6.19	17.72±5.33	20.82±7.66	28.98±9.54
	MVP	62.33±3.06	48.32±11.42	50.23±11.27	41.55±1.98	29.29±5.03	62.17±6.10	28.73±3.13	23.62±9.51	62.20±5.81
	MISA	65.30±2.28	67.43±6.75	17.34±6.20	40.94±1.22	36.16±1.58	35.62±11.69	18.62±3.36	23.66±2.01	21.18±10.00
	FlyPrompt	75.58±1.70	79.36±3.47	8.38±1.52	57.75±2.12	54.39±1.29	23.40±3.20	28.86±5.84	36.79±7.58	18.85±4.15
	MIST	**83.64**±2.20	**86.91**±0.35	**4.36**±0.60	**66.53**±1.15	**65.91**±1.44	**8.39**±0.90	**50.01**±3.77	**60.76**±10.38	**11.29**±11.24
iBOT-1K	L2P	53.17±7.08	62.28±8.19	16.47±7.89	38.29±2.65	39.86±0.95	33.01±8.96	19.20±2.21	31.21±5.24	21.26±11.09
	DualPrompt	52.39±3.01	53.56±6.10	20.45±3.88	45.76±1.63	39.19±0.65	35.14±7.72	29.32±3.15	30.53±5.33	32.70±9.00
	CODA-P	59.29±4.03	61.30±6.73	22.72±6.32	49.56±1.57	42.64±2.78	40.23±6.19	27.57±2.83	33.61±5.52	27.71±9.65
	MVP	57.52±3.62	44.08±12.42	53.47±12.87	44.76±2.23	34.93±4.48	56.26±5.49	33.81±2.35	26.32±9.97	62.61±6.88
	MISA	54.31±2.91	55.89±5.10	19.16±4.04	43.91±3.95	40.09±1.24	33.52±10.12	27.76±2.69	33.74±5.11	29.35±6.60
	FlyPrompt	70.14±1.76	74.84±4.26	11.52±3.71	61.50±1.66	57.18±1.36	20.89±3.75	38.75±5.72	45.00±4.19	21.62±3.82
	MIST	**79.38**±2.67	**84.60**±0.49	**4.88**±1.24	**70.14**±1.23	**70.40**±1.48	**7.44**±1.10	**60.96**±3.10	**68.36**±10.27	**9.20**±12.16
DINO-1K	L2P	47.98±7.33	59.13±6.32	14.82±6.81	35.81±1.37	36.58±1.31	33.00±7.22	21.18±2.01	32.47±6.10	24.23±9.93
	DualPrompt	52.12±4.01	55.71±6.11	18.69±5.27	43.03±1.12	35.40±1.40	37.89±7.65	27.80±4.21	29.49±4.24	31.35±10.61
	CODA-P	54.69±4.49	58.91±5.43	19.43±5.83	45.16±2.05	38.23±2.02	41.26±6.20	29.22±2.97	31.85±7.47	32.69±10.65
	MVP	53.64±3.91	41.02±12.09	54.70±12.47	41.78±2.15	32.00±4.22	56.54±6.00	33.44±3.43	26.02±10.29	64.21±7.82
	MISA	52.03±3.07	55.98±4.26	17.76±4.28	41.26±3.25	37.50±1.62	34.38±10.60	27.13±3.31	33.08±4.10	25.77±10.77
	FlyPrompt	65.92±2.74	72.66±4.52	9.04±1.73	57.29±2.40	54.72±1.89	20.65±3.08	37.38±5.86	44.66±2.35	19.47±4.84
	MIST	**76.32**±2.71	**82.28**±0.56	**5.14**±0.97	**65.24**±1.45	**65.73**±1.23	**7.21**±1.26	**61.63**±3.23	**69.04**±8.57	**9.58**±11.68
MoCo v3-1K	L2P	28.17±7.08	39.07±11.31	19.68±15.19	17.43±1.71	16.27±5.43	43.59±4.64	12.42±2.31	20.00±7.36	26.38±17.63
	DualPrompt	53.33±4.65	58.20±7.73	14.79±4.09	36.69±1.74	30.24±1.94	35.07±5.93	19.88±3.35	21.93±4.30	28.42±9.47
	CODA-P	53.47±3.42	58.55±7.19	16.97±6.15	39.89±2.71	31.72±4.86	43.27±2.27	20.09±2.52	24.10±6.48	30.90±10.77
	MVP	54.33±4.56	40.84±14.21	55.42±15.00	36.45±2.35	26.37±6.04	55.87±4.26	28.48±3.34	23.56±9.78	62.97±5.67
	MISA	57.00±6.06	62.18±3.94	16.19±5.26	38.85±4.27	33.47±0.95	34.42±8.60	25.02±4.39	27.68±4.35	35.50±7.97
	FlyPrompt	64.12±5.18	71.51±8.48	7.69±2.42	52.32±1.50	49.06±1.35	21.27±2.89	27.92±4.53	33.32±3.58	20.69±3.58
	MIST	**76.22**±2.88	**81.88**±0.52	**6.35**±1.17	**58.74**±1.41	**59.73**±1.51	**6.16**±1.06	**49.50**±5.15	**55.04**±1.81	**13.08**±17.48

where $\bar{f}_{\mathcal{B}} = |\mathcal{B}|^{-1} \sum_{\mathbf{x} \in \mathcal{B}} [1 - \nu(\mathbf{x})]$ modulates consolidation according to statistical familiarity, and $\lambda_{\text{con}}, \lambda_{\text{geo}} \geq 0$ control the two regularizers. Prototype updates and the complete optimization procedure are detailed in Appendix E.2.

4 EXPERIMENT

4.1 EXPERIMENT SETUP

Benchmarks. We evaluate MIST under the Si-Blurry GCL setting (Moon et al., 2023; Kang et al., 2025; Yan et al., 2026) on CIFAR-100 (Krizhevsky et al., 2009), ImageNet-R (Hendrycks et al., 2021), and CUB-200-2011 (Wah et al., 2011), containing 60K, 30K, and 12K images across 100, 200, and 200 classes, respectively. Unless otherwise specified, we use $r_D = 50\%$ and $r_B = 10\%$ with single-pass training. We vary the number of tasks $T \in \{5, 10, 15, 20\}$ for sensitivity analysis. Evaluation metrics include average anytime accuracy (A_{auc}), final accuracy (A_{last}) (Koh et al., 2021), final forgetting (F_{last}), and backward transfer (BWT) (Lin et al., 2022). Results are reported as mean ± standard deviation over five random seeds. Metric definitions appear in Appendix C; ratio sensitivity results are in Appendices G.3 and G.4.

Baselines and Pretrained Models. We compare MIST with three categories of CL and GCL methods: (i) reference baselines, including sequential fine-tuning (Seq FT), its slow-learning-rate variant (Seq FT w/ SL) (Zhang et al., 2023), and linear probing; (ii) prompt-based CL methods, including L2P (Wang et al., 2022d), DualPrompt (Wang et al., 2022c), and CODA-Prompt (Smith

Table 2: Ablation study on CUB-200 and ImageNet-R across different PTMs. CSM, CIT, GC, and APF denote continuous statistical measurement (Sec.3.1), conditional isometric transport (Sec.3.2), geometry consolidation (Sec.3.2), and adaptive prediction fusion (Sec.3.3), respectively.

PTM	Components				CUB-200			ImageNet-R		
	CSM	CIT	GC	APF	A_{auc} (%)	A_{last} (%)	F_{last} (%)	A_{auc} (%)	A_{last} (%)	F_{last} (%)
Sup-21K	✓	×	×	×	72.00 ± 3.29	71.70 ± 0.72	13.23 ± 0.98	39.86 ± 0.90	43.61 ± 0.83	8.24 ± 1.97
	✓	✓	×	×	72.58 ± 2.74	73.67 ± 0.95	13.00 ± 0.88	44.24 ± 1.08	48.92 ± 1.20	9.84 ± 1.36
	✓	✓	✓	×	76.22 ± 2.50	79.84 ± 2.94	8.46 ± 2.89	46.33 ± 1.03	49.53 ± 1.16	9.15 ± 1.75
	✓	✓	✓	✓	**76.98** ± 2.02	**80.58** ± 1.54	**5.32** ± 2.00	**51.96** ± 1.48	**50.68** ± 1.00	**7.59** ± 1.19
MoCo v3-1K	✓	×	×	×	44.14 ± 3.94	52.96 ± 2.14	**10.46** ± 5.06	53.89 ± 1.70	57.10 ± 2.05	11.38 ± 1.51
	✓	✓	×	×	46.79 ± 3.97	53.77 ± 2.83	11.11 ± 5.27	53.91 ± 1.29	57.81 ± 0.89	12.85 ± 1.84
	✓	✓	✓	×	48.55 ± 2.97	53.28 ± 1.43	15.92 ± 15.27	56.10 ± 1.41	57.15 ± 4.43	12.05 ± 3.75
	✓	✓	✓	✓	**49.50** ± 5.15	**55.04** ± 1.81	13.08 ± 17.48	**58.74** ± 1.41	**59.73** ± 1.51	**6.16** ± 1.06

et al., 2023); and (iii) replay-free GCL methods, including MVP (Moon et al., 2023), MISA (Kang et al., 2025), and FlyPrompt (Yan et al., 2026). MIST and all prompt-based baselines are evaluated with six ViT-B/16 pretrained models: two supervised configurations (Sup-21K (Ridnik et al., 2021) and Sup-21K/1K (Dosovitskiy et al., 2020), the latter fine-tuned on ImageNet-1K) and four self-supervised configurations (iBOT-21K, iBOT-1K (Zhou et al., 2021), DINO-1K (Caron et al., 2021), and MoCo v3-1K (Chen et al., 2021)). Within each configuration, all methods use the same pretrained initialization.

Implementation. We use a frozen ViT-B/16 backbone and train MIST with Adam (1×10^{-3} learning rate, batch size 64), performing two optimization steps per incoming mini-batch in a single-pass stream. The shared prompt contains $P = 20$ tokens of dimension $d_p = 32$, inserted into the first five Transformer blocks. We set $d_m = 64$, $R_g = 8$, and $E_{\max} = 5$, with two EMA classifiers per interval. Further details are provided in Appendix B.

4.2 EXPERIMENT RESULTS

Overall Performance. Tab.1 summarizes the performance of MIST across three GCL benchmarks and six pretrained-model configurations. Overall, MIST delivers consistently strong results across both supervised and self-supervised PTMs. Compared with the strongest competing methods, it provides improvements of up to 12.10, 9.14, and 24.25 percentage points in A_{auc}, and 10.37, 14.50, and 24.38 points in A_{last} on CIFAR-100, ImageNet-R, and CUB-200, respectively. The gains become particularly pronounced under self-supervised pretraining, where existing prompt-based approaches exhibit substantial degradation, especially on the fine-grained CUB-200 benchmark. In contrast, MIST maintains substantially stronger continual performance across iBOT, DINO, and MoCo v3 representations, indicating improved robustness to variations in pretrained feature spaces. Moreover, MIST consistently achieves lower F_{last} across the evaluated configurations, showing that the improved online and final accuracy is accompanied by stronger retention rather than increased forgetting. The only accuracy degradation occurs on ImageNet-R under Sup-21K; nevertheless, MIST reduces F_{last} from 13.42% to 7.59%, while recovering a clear accuracy advantage under all other pretrained configurations. Additional comparisons with offline CL methods are provided in Appendix G.1.

Ablation Study. Table 2 evaluates progressive additions with CSM held fixed. CIT improves mean anytime and final accuracy, supporting input-dependent adaptation within a shared prompt structure. GC further improves anytime accuracy, but its effects on forgetting vary across configurations: geometric constraints alone do not ensure retention. Adding APF improves both accuracy metrics and reduces forgetting relative to the preceding variant. The full model nevertheless does not minimize forgetting in every setting. These patterns support complementary representation and prediction mechanisms for reusing adaptation knowledge while responding to changing class mixtures; this design does not isolate the contribution of CSM.

Sensitivity to the Number of Tasks. Fig. 3 examines the sensitivity of GCL methods to the number of tasks under the self-supervised iBOT-21K backbone. MIST consistently achieves the highest

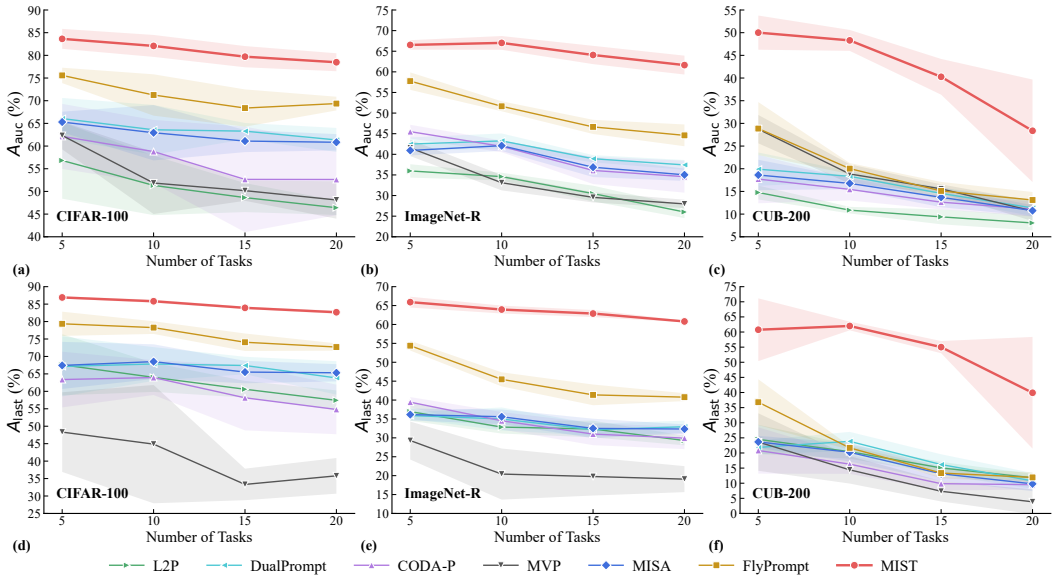

Figure 3: GCL performance with varying task numbers under iBOT-21K. (a)–(c) show A_{auc} and (d)–(f) show A_{last} on CIFAR-100, ImageNet-R, and CUB-200, respectively. Lines and shaded regions indicate the mean and standard deviation over five seeds.

A_{auc} and A_{last} across all task numbers on the three benchmarks. Its performance decreases gradually on CIFAR-100 and ImageNet-R as the task partition becomes finer, while a larger decline is observed on CUB-200. Nevertheless, MIST preserves a clear advantage over competing methods, particularly on CUB-200. These results indicate that the shared latent prompt and measurement-conditioned isometric transport remain effective under varying task granularities, although increasingly fragmented streams introduce greater difficulty for fine-grained recognition. Additional numerical results are provided in Appendix G.2.

Performance and Parameter Efficiency. Tab.3 compares continual learning performance and parameter efficiency. MIST achieves the highest mean A_{auc} and BWT with the fewest trainable parameters. Compared with FlyPrompt, it reduces trainable parameters by 56.5% while maintaining comparable accuracy and higher BWT. The results demonstrate that MIST combines parameter-efficient adaptation with positive backward transfer without compromising overall accuracy.

Table 3: Performance, backward transfer, and parameter efficiency on CIFAR-100 under Sup-21K.

Method	A_{auc} (%)	BWT (%)	Total (M)	Trainable (M)
L2P	76.23 ± 2.73	0.10 ± 2.65	**86.01**	0.22
DualPrompt	76.04 ± 3.32	-2.93 ± 2.42	86.35	0.55
CODA-P	79.13 ± 3.06	-0.83 ± 2.17	86.72	0.92
MVP	67.74 ± 4.96	-18.09 ± 3.24	86.12	0.32
MISA	80.35 ± 2.39	-1.76 ± 2.28	86.37	0.58
FlyPrompt	82.65 ± 2.56	4.35 ± 1.19	87.08	0.46
MIST	$\mathbf{82.98 \pm 1.59}$	$\mathbf{5.98 \pm 1.24}$	86.85	**0.20**

5 Conclusion

We presented MIST, which formulates continual adaptation as conditional variation of a shared prompt structure to jointly address knowledge reuse and adaptation to non-stationary distributions. A frozen statistical reference supports consistent historical associations, while conditional isometric transport preserves latent token geometry at a fixed shared configuration and retains flexibility for input-dependent adaptation. The same associations connect prompt generation with classifier aggregation, and novelty and confidence regulate the contributions of historical and online plastic predictions. Theoretical analysis and empirical evaluation support the complementary roles of structured adaptation and adaptive decision making. MIST expresses knowledge reuse through the conditional use of shared structure and coordinates historical information with current evidence at the prediction level, providing a structured approach to balancing stability and plasticity in continual learning.

Reproducibility Statement

To ensure the reproducibility of our work, we have provided comprehensive details about the experimental setup in Section 4.1, including datasets used, baseline methods, implementation details and evaluation metrics. More details about the implementation can be found in Appendix B. All code, models, and configuration files required to replicate our results are made publicly available in a repository.

Large Language Models Assistance

Large language models were used to assist with manuscript editing, technical discussions, and the presentation of mathematical arguments. The authors independently verified the final content and take full responsibility for the work.

References

Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. In *CVPR*, pp. 11246–11255, 2019a.

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *NeurIPS*, 32, 2019b.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *NeurIPS*, 33:15920–15930, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9630–9640, 2021.

Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co 2 l: Contrastive continual learning. In *ICCV*, pp. 9496–9505, 2021.

Huancheng Chen, Jingtao Li, Weiming Zhuang, Chen Chen, and Lingjuan Lyu. Replay-free continual low-rank adaptation with dynamic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3161–3170, 2026.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pp. 9620–9629, 2021.

Quan Cheng, Yuanyu Wan, Lingyu Wu, Chenping Hou, and Lijun Zhang. Continuous subspace optimization for continual learning. *NeurIPS*, 38:15376–15398, 2025.

Andrew M Childs and Wim Van Dam. Quantum algorithms for algebraic problems. *Reviews of Modern Physics*, 82(1):1–52, 2010.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *PAMI*, 2021.

Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Loss of plasticity in deep continual learning. *Nature*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *ECCV*, pp. 86–102, 2020.

Yonina C Eldar and G David Forney. Optimal tight frames and quantum measurement. *IEEE Transactions on Information Theory*, 48(3):599–610, 2002.

Alex Gomez-Villa, Dipam Goswami, Kai Wang, Andrew D Bagdanov, Bartlomiej Twardowski, and Joost Van De Weijer. Exemplar-free continual representation learning via learnable drift compensation. In *ECCV*, pp. 473–490, 2024.

Dipam Goswami, Yuyang Liu, Bartłomiej Twardowski, and Joost Van De Weijer. Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. *NeurIPS*, 36:6582–6595, 2023.

Paul Hausladen and William K Wootters. A 'pretty good'measurement for distinguishing quantum states. *Journal of Modern Optics*, 41(12):2385–2390, 1994.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pp. 8320–8329, 2021.

Chenxing Hong, Yan Jin, Zhiqi Kang, Yizhou Chen, Mengke Li, Yang Lu, and Hanzi Wang. Dynamically anchored prompting for task-imbalanced continual learning. In *IJCAI*, 2024.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pp. 709–727, 2022.

Zhiqi Kang, Enrico Fini, Moin Nabi, Elisa Ricci, and Karteek Alahari. A soft nearest-neighbor framework for continual semi-supervised learning. In *ICCV*, pp. 11834–11843, 2023.

Zhiqi Kang, Liyuan Wang, Xingxing Zhang, and Karteek Alahari. Advancing prompt-based methods for replay-independent general continual learning. In *ICLR*, pp. 50707–50725, 2025.

Hyunseo Koh, Dahyun Kim, Jung-Woo Ha, and Jonghyun Choi. Online continual learning on class incremental blurry task configuration with anytime inference. *arXiv preprint arXiv:2110.10031*, 2021.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Minh Le, An Nguyen, Huy Nguyen, Trang Nguyen, Trang Pham, Linh Van Ngo, and Nhat Ho. Mixture of experts meets prompt-based continual learning. *NeurIPS*, 37:119025–119062, 2024.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 3045–3059, 2021.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing*, pp. 4582–4597, 2021.

Zhuowei Li, Long Zhao, Zizhao Zhang, Han Zhang, Di Liu, Ting Liu, and Dimitris N Metaxas. Steering prototypes with prompt-tuning for rehearsal-free continual learning. In *WACV*, pp. 2511–2521, 2024.

Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. Beyond not-forgetting: Continual learning with backward knowledge transfer. *NeurIPS*, 35:16165–16177, 2022.

Fei Mi, Lingjing Kong, Tao Lin, Kaicheng Yu, and Boi Faltings. Generalized class incremental learning. In *CVPR*, pp. 970–974, 2020.

Maher Moakher. Means and averaging in the group of rotations. *SIAM journal on matrix analysis and applications*, 24(1):1–16, 2002.

Jun-Yeong Moon, Keon-Hee Park, Jung Uk Kim, and Gyeong-Moon Park. Online class incremental learning on stochastic blurry task boundary via mask and visual prompt tuning. In *ICCV*, pp. 11697–11707, 2023.

Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jahnichen, and Moin Nabi. Learning to remember: A synaptic plasticity driven framework for continual learning. In *CVPR*, pp. 11321–11329, 2019.

Seoyoung Park, Haemin Lee, and Hankook Lee. Is prompt selection necessary for task-free online continual learning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 7883–7892, 2026.

Jingyang Qiao, Zhizhong Zhang, Xin Tan, Yanyun Qu, Wensheng Zhang, Zhi Han, and Yuan Xie. Gradient projection for continual parameter-efficient tuning. *PAMI*, 47(10):9316–9329, 2025.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *NeurIPS*, 36:79320–79362, 2023.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *ICML*, pp. 4548–4557, 2018.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *CVPR*, pp. 11909–11919, 2023.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning. In *ECCV*, pp. 398–414, 2022a.

Huaijie Wang, De Cheng, Lingfeng He, Yan Li, Jie Li, Nannan Wang, and Xinbo Gao. Ekpc: Elastic knowledge preservation and compensation for class-incremental learning: H. wang et al. *IJCV*, 134(5):238, 2026.

Liyuan Wang, Jingyi Xie, Xingxing Zhang, Mingyi Huang, Hang Su, and Jun Zhu. Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. *NeurIPS*, 36:69054–69076, 2023.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *PAMI*, 2024.

Liyuan Wang, Jingyi Xie, Xingxing Zhang, Hang Su, and Jun Zhu. Hide-pet: continual learning via hierarchical decomposition of parameter-efficient tuning. *PAMI*, 47(8):6687–6702, 2025.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-Prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *NeurIPS*, 35:5682–5695, 2022b.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *ECCV*, pp. 631–648, 2022c.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, pp. 139–149, 2022d.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *NeurIPS*, 34:4738–4750, 2021.

Yichen Wu, Hongming Piao, Long-Kai Huang, Renzhen Wang, Wanhua Li, Hanspeter Pfister, Deyu Meng, Kede Ma, and Ying Wei. Sd-lora: Scalable decoupled low-rank adaptation for class incremental learning. In *ICLR*, pp. 58337–58353, 2025.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *CVPR*, pp. 374–382, 2019.

Hongwei Yan, Guanglong Sun, Kanglei Zhou, Qian Li, Liyuan Wang, and Yi Zhong. Flyprompt: Brain-inspired random-expanded routing with temporal-ensemble experts for general continual learning. In *ICLR*, 2026.

Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *CVPR*, pp. 6982–6991, 2020.

Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. In *ICCV*, 2023.

Da-Wei Zhou, Hai-Long Sun, Jingyi Ning, Han-Jia Ye, and De-Chuan Zhan. Continual learning with pre-trained models: A survey. *arXiv preprint arXiv:2401.16386*, 2024.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

A SI-BLURRY DATA PROTOCOL

We adopt the stochastic incremental blurry (Si-Blurry) protocol (Moon et al., 2023), following its use in recent GCL studies (Kang et al., 2025; Yan et al., 2026). The protocol constructs a sequential stream $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ from a class set $\mathcal{C}$, where N denotes the total number of training samples and $C = |\mathcal{C}|$. The stream comprises T benchmark segments, indexed by $t \in \{1, \ldots, T\}$.

Class allocation. The classes are randomly separated into disjoint classes $\mathcal{C}^D$ and blurry classes $\mathcal{C}^B$:

$$\mathcal{C} = \mathcal{C}^D \cup \mathcal{C}^B, \quad \mathcal{C}^D \cap \mathcal{C}^B = \varnothing, \quad r_D = \frac{|\mathcal{C}^D|}{|\mathcal{C}|}. \tag{9}$$

Each subset is then randomly partitioned among the T segments, producing initial class allocations $\{\mathcal{C}_t^D\}_{t=1}^T$ and $\{\mathcal{C}_t^B\}_{t=1}^T$. These allocations do not require equal numbers of classes across segments. Samples from each disjoint class remain within its assigned segment, whereas samples from blurry classes can be distributed across multiple segments.

Sample redistribution. The blurry-class sample pool is $\mathcal{D}^B = \{(\mathbf{x}_i, y_i) \in \mathcal{D} : y_i \in \mathcal{C}^B\}$. A randomly selected subset $\mathcal{D}^{B,\mathrm{mix}} \subseteq \mathcal{D}^B$ is redistributed across segments. The blurry sample ratio specifies its proportion within this pool:

$$r_B = \frac{|\mathcal{D}^{B,\mathrm{mix}}|}{|\mathcal{D}^B|}, \qquad |\mathcal{D}^B| > 0, \tag{10}$$

up to integer rounding. The remaining blurry-class samples retain their initial segment assignments. Consequently, the samples in segment t comprise

$$\mathcal{D}_t = \mathcal{D}_t^D \cup \mathcal{D}_t^{B,\mathrm{keep}} \cup \mathcal{D}_t^{B,\mathrm{mix}}, \tag{11}$$

where the three terms contain samples from the assigned disjoint classes, retained samples from the assigned blurry classes, and redistributed blurry-class samples, respectively. Each training example is assigned to one segment; class recurrence arises from different examples sharing the same label. This construction permits varying class counts, overlapping class exposure, and unequal class frequencies across segments. At $r_D = 0$, all classes belong to the blurry subset; at $r_D = 1$, class allocations are fully disjoint and r_B has no effect.

Online learning and interval indexing. Samples are shuffled within each segment, and the segments are concatenated in temporal order. Under our replay-free evaluation, the resulting stream is traversed once without revisiting previous mini-batches. Benchmark segment identities are used for stream construction and evaluation and are not provided to MIST. They are distinct from the internal sample-count intervals indexed by e, with $E \leq E_{\max}$ instantiated intervals. These internal intervals organize the model's statistical memories and adaptation states and need not coincide with benchmark segments. Unless otherwise specified, we use $r_D = 50\%$ and $r_B = 10\%$.

B IMPLEMENTATION DETAILS

MIST uses a frozen ViT-B/16 with embedding dimension $D = 768$. Images are processed at 224×224 resolution with AutoAugment during training. The experimental configuration in the main text uses Adam with learning rate 10^{-3}, batch size 64, and $K = 2$ optimization steps on each incoming mini-batch. The stream is traversed once; these local steps do not revisit earlier mini-batches. Unless otherwise specified, the Si-Blurry stream uses 5 benchmark segments, $r_D = 50\%$, and $r_B = 10\%$.

The number of benchmark segments T and the number of internal intervals $E_{\max}$ have different roles. The former specifies stream construction and evaluation; the latter allocates statistical memories and adaptation states. The main experiments use $E_{\max} = 5$, independently of the benchmark granularity unless a configuration explicitly changes it. For a stream of N samples, the prescribed boundaries are $b_k = \lfloor kN/E_{\max} \rfloor$, $k = 1, \ldots, E_{\max} - 1$. The training loop checks the sample-count scheduler after processing an incoming mini-batch; newly activated states apply to subsequent batches. No benchmark task identity is supplied to the model.

Table 4: MIST configuration. The cosine scale is initialized at 16 and subsequently learned. Internal interval capacity is distinct from the number of benchmark segments.

Quantity	Symbol	Setting
Prompted blocks, using one-based indices	$\mathcal{L}$	$\{1, \ldots, 5\}$
Latent tokens per prompted block	P	20
Latent prompt dimension	d_p	32
Measurement dimension	d_m	64
Givens planes per block	R_g	8
Internal interval capacity	$E_{\max}$	5
EMA heads per interval	M	2
EMA decay rates	$\{\alpha_m\}$	$\{0.9, 0.99\}$
Isotropic reference prior	π_0	0.08
Historical mixing during training	β	0.2
Measurement stabilization	ϵ	10^{-5}
Prompt-moment denominator floor	ϵ_p	10^{-8}
Gate center and temperature	(τ_ν, T_ν)	$(0.08, 0.04)$
Confidence-gate slope	κ	8
Initial cosine scale; permitted range	s	16; $[1, 100]$
Shared parameter initialization std.	—	0.02
New-interval angle perturbation std.	—	0.005

C Evaluation metrics

At evaluation checkpoint q, $\mathcal{T}_q$ contains held-out examples from classes observed so far. Its sample-weighted accuracy is $a_q = |\mathcal{T}_q|^{-1} \sum_{(x,y)\in\mathcal{T}_q} \mathbb{I}(\hat{y}_q(x) = y)$. Average anytime accuracy and final accuracy are

$$A_{\text{auc}} = \frac{1}{Q}\sum_{q=1}^{Q} a_q, \qquad A_{\text{last}} = \frac{1}{|\mathcal{T}_{\text{final}}|} \sum_{(x,y)\in\mathcal{T}_{\text{final}}} \mathbb{I}(\hat{y}_{\text{final}}(x) = y). \tag{12}$$

The former evaluates performance throughout learning; the latter measures its endpoint. The point-average convention follows anytime evaluation (Koh et al., 2021); it approximates normalized area under the accuracy curve for uniformly spaced checkpoints. All methods must share the same checkpoint schedule and evaluation label space.

For benchmark stages $t = 1, \ldots, T$, $R_{t,i}$ is accuracy after stage t on the fixed held-out cohort associated with stage i. Following the stage-wise convention used here, final forgetting and backward transfer are

$$F_{\text{last}} = \frac{1}{T}\sum_{i=1}^{T} \left(\max_{j=i,\ldots,T} R_{j,i} - R_{T,i} \right), \qquad \text{BWT} = \frac{1}{T-1}\sum_{i=1}^{T-1} (R_{T,i} - R_{i,i}). \tag{13}$$

Forgetting measures decline from each cohort's recorded peak, whereas BWT measures change from its post-acquisition performance (Lin et al., 2022). Including the final checkpoint in the maximum makes $F_{\text{last}} \geq 0$; its last term is zero. BWT is defined for $T > 1$ and can have either sign. Its different reference and denominator mean that it is not simply $-F_{\text{last}}$. Under class recurrence, both metrics can reflect renewed class exposure as well as interference or transfer. Stage labels are used only for evaluation, independently of internal intervals. Results report mean $\pm$ standard deviation over five seeds in percent; comparisons of means do not constitute significance tests.

D MATHEMATICAL PROPERTIES

D.1 POSITIVE MEASUREMENT AND STATISTICAL ASSOCIATION

For an evaluation-transformed batch represented by the normalized query matrix Q, the active interval accumulates $H_e \leftarrow H_e + Q^\top Q$ and $N_e \leftarrow N_e + |Q|$, where $|Q|$ denotes the number of queries. The corresponding second-order statistic and its normalized form are

$$\begin{aligned} C_e &= \frac{H_e}{\max(N_e, 1)}, \\ \rho_e &= \frac{C_e + \epsilon I}{\mathrm{Tr}(C_e + \epsilon I)}. \end{aligned} \tag{14}$$

The fixed orthonormal projection is initialized by reduced QR decomposition of a Gaussian matrix. Statistics are collected once per incoming batch, after its local optimization steps, using evaluation preprocessing.

The isotropic reference is $\rho_0 = I/d_m$, with prior $\pi_0 \in (0, 1)$. Historical intervals receive the smoothed priors

$$\pi_e = (1 - \pi_0)\frac{N_e + 1}{\sum_{j=1}^{E}(N_j + 1)}, \qquad e = 1, \ldots, E. \tag{15}$$

For the mixture $S = \sum_{e=0}^{E} \pi_e \rho_e$, the pretty-good measurement construction (Hausladen & Wootters, 1994; Childs & Van Dam, 2010; Eldar & Forney, 2002) defines

$$\begin{aligned} M_e &= S^{-1/2}\pi_e\rho_e S^{-1/2}, && e = 0, \ldots, E, \\ a_e &= qM_e q^\top, && e = 0, \ldots, E, \\ \nu &= a_0, \\ r_e &= \frac{a_e}{1 - a_0}, && e = 1, \ldots, E. \end{aligned} \tag{16}$$

Proposition D.1 (Positivity and completeness)**.** *For $\epsilon > 0$, $\pi_0 \in (0, 1)$, $E \geq 1$, and a unit query q, the measurement operators satisfy $M_e \succeq 0$ and $\sum_{e=0}^{E} M_e = I$. Consequently, their responses satisfy $a_e \geq 0$ and $\sum_{e=0}^{E} a_e = 1$, and the historical association vector r is well defined on the interval simplex.*

Derivation. Each ρ_e is positive definite with unit trace, including an empty interval whose statistic is I/d_m. The isotropic reference ensures $S \succeq (\pi_0/d_m)I \succ 0$, so its inverse square root exists. Congruence by $S^{-1/2}$ preserves positivity, while summation gives

$$\begin{aligned} \sum_{e=0}^{E} M_e &= S^{-1/2}\left(\sum_{e=0}^{E}\pi_e\rho_e\right)S^{-1/2} \\ &= S^{-1/2}SS^{-1/2} \\ &= I. \end{aligned} \tag{17}$$

Applying this identity to a unit query establishes response normalization. Positive interval priors and positive-definite statistics further ensure $\sum_{e=1}^{E} a_e = 1 - a_0 > 0$. The interval normalization therefore produces nonnegative weights summing to one.

Interpretation of the reference response. For the eigendecomposition $S = \sum_k \lambda_k v_k v_k^\top$, the reference response satisfies

$$\begin{aligned} \nu &= \frac{\pi_0}{d_m} q S^{-1} q^\top \\ &= \frac{\pi_0}{d_m}\sum_{k=1}^{d_m}\frac{(qv_k)^2}{\lambda_k}. \end{aligned} \tag{18}$$

Directions with smaller second-order mass in the accumulated mixture receive greater weight. Thus, ν measures relative statistical novelty with respect to historical structures, rather than a calibrated probability of an unseen class.

Numerical implementation. For numerical stability, eigenvalues are lower-bounded by ϵ, negative round-off responses are clipped to zero, and the complete response vector is normalized before its interval components are renormalized. With $\pi_0 = 0.08$ and $d_m = 64$, the lower bound $\lambda_{\min}(S) \geq 0.00125$ exceeds $\epsilon = 10^{-5}$, so eigenvalue clipping is inactive in exact arithmetic. The identities above assume unit queries and the exact inverse square root; numerical safeguards address near-degenerate queries and finite-precision deviations.

D.2 Conditional Isometry and Sensitivity

For one prompted layer, suppressing its index, the shared latent configuration Z_0 is conditioned through

$$\begin{aligned} \bar{\theta}_r(w) &= \sum_{e=1}^{E} w_e \theta_{e,r}, \\ U(w) &= \prod_{r=1}^{R_g} G_r\big(\bar{\theta}_r(w)\big), \\ Z(w) &= Z_0 U(w), \\ P(w) &= Z(w)W + B. \end{aligned} \tag{19}$$

The Givens rotations follow a fixed plane order, and B is independent of the conditioning vector.

Proposition D.2 (Latent geometry and conditional variation)**.** *For fixed shared parameters Z_0, W and interval coordinates, the transported prompts satisfy*

$$\begin{aligned} Z(w)Z(w)^\top &= Z_0 Z_0^\top, \\ \|P(w) - P(w')\|_F &\leq \|Z_0\|_F \|W\|_2 \min\{2, d_\theta(w, w')\}, \end{aligned} \tag{20}$$

where

$$d_\theta(w, w') = \sum_{r=1}^{R_g} \left| \sum_{e=1}^{E} (w_e - w'_e)\theta_{e,r} \right|. \tag{21}$$

For simplex weights and any reference coordinate vector $\theta_\star$,

$$d_\theta(w, w') \leq \max_e \|\theta_e - \theta_\star\|_1 \|w - w'\|_1. \tag{22}$$

Derivation. Every Givens factor is orthogonal, hence $U(w)U(w)^\top = I$. Substitution gives

$$Z(w)Z(w)^\top = Z_0 U(w) U(w)^\top Z_0^\top = Z_0 Z_0^\top. \tag{23}$$

For a single rotation plane,

$$\|G_r(a) - G_r(b)\|_2 = 2 \left| \sin\left(\frac{a-b}{2}\right) \right| \leq |a - b|. \tag{24}$$

A telescoping expansion of the ordered products, whose remaining factors have unit operator norm, yields $\|U(w) - U(w')\|_2 \leq d_\theta(w, w')$. Orthogonality also gives $\|U(w) - U(w')\|_2 \leq 2$. Therefore,

$$\|P(w) - P(w')\|_F = \big\|Z_0\big(U(w) - U(w')\big)W\big\|_F \leq \|Z_0\|_F \|U(w) - U(w')\|_2 \|W\|_2, \tag{25}$$

which proves the sensitivity bound. Since $\sum_e (w_e - w'_e) = 0$, subtracting $\theta_\star$ does not change the angle difference. The triangle inequality then establishes equation 22.

The Gram identity preserves latent token norms, inner products, and distances across conditions at a fixed model state. It does not assert invariance across updates to Z_0 or after a general learned map W. The sensitivity bound separates conditional variation from changes caused by optimization. Mixing angles before constructing rotations retains orthogonality without averaging prompt tensors or rotation matrices.

D.3 GEOMETRY CONSOLIDATION

The token Gram matrix $ZZ^\top$ is invariant under right-orthogonal transport, whereas the coordinate moment $Z^\top Z$ changes orientation. Consolidation regularizes the latter. For a historical interval $j < e$, its reference is

$$\begin{aligned} Z_j^l &= Z_0^l \prod_{r=1}^{R_g} G_r(\theta_{j,r}^l), \\ \Gamma_j^l &= \mathrm{sg}\left(\frac{(Z_j^l)^\top Z_j^l}{\max\left(\|Z_j^l\|_F^2, \epsilon_p\right)} \right), \end{aligned} \tag{26}$$

where sg denotes stop-gradient. These references are recomputed from the current shared prompt and historical coordinates, rather than stored as snapshots of earlier shared prompts.

The active coordinates define $A^l = \sum_{r=1}^{R_g} \theta_{e,r}^l B_r$, where $(B_r)_{i_r j_r} = 1$, $(B_r)_{j_r i_r} = -1$, and all other entries vanish. Writing $[A, \Gamma] = A\Gamma - \Gamma A$, the two penalties are

$$\begin{aligned} \mathcal{L}_{\mathrm{con}} &= \frac{\sum_{l \in \mathcal{L}} \sum_{j<e} \left\| [A^l, \Gamma_j^l] \right\|_F^2}{|\mathcal{L}|(e-1)d_p^2}, \\ \mathcal{L}_{\mathrm{geo}} &= \frac{\sum_{l \in \mathcal{L}} \left\| \theta_e^l \right\|_2^2}{|\mathcal{L}| R_g}. \end{aligned} \tag{27}$$

The first penalty is defined as zero for $e = 1$. The factor d_p^2 normalizes consolidation by the number of matrix entries.

Proposition D.3 (Local control of the coordinate moment)**.** *For a fixed symmetric matrix Γ, coordinates θ, and $U(t) = \prod_{r=1}^{R_g} G_r(t\theta_r)$, define $A = \sum_{r=1}^{R_g} \theta_r B_r$. Then*

$$\frac{d}{dt}\left(U(t)^\top \Gamma U(t) \right)\Big|_{t=0} = [A, \Gamma], \tag{28}$$

and the finite variation satisfies

$$\left\| U(t)^\top \Gamma U(t) - \Gamma \right\|_F \le |t| \, \|[A, \Gamma]\|_F + 2t^2 \|\theta\|_1^2 \|\Gamma\|_F. \tag{29}$$

Derivation. The sign convention gives $G_r(t\theta_r) = \exp(-t\theta_r B_r)$, so $U'(0) = -A$. Since $A^\top = -A$, differentiating the transported moment at the origin gives $A\Gamma - \Gamma A$.

The product rule and orthogonality imply

$$\begin{aligned} \|U'(t)\|_2 &\le \|\theta\|_1, \\ \|U''(t)\|_2 &\le \|\theta\|_1^2. \end{aligned} \tag{30}$$

Writing $F(t) = U(t)^\top \Gamma U(t)$, its second derivative is

$$F''(t) = U''(t)^\top \Gamma U(t) + 2U'(t)^\top \Gamma U'(t) + U(t)^\top \Gamma U''(t). \tag{31}$$

Consequently, $\|F''(t)\|_F \le 4\|\theta\|_1^2 \|\Gamma\|_F$. Taylor's theorem with integral remainder establishes equation 29.

Since $\|\theta\|_1^2 \le R_g \|\theta\|_2^2$, the angle penalty also controls the remainder bound. For $\Gamma = V \, \mathrm{diag}(\lambda_i) V^\top$, the consolidation penalty admits the spectral expression

$$\|[A, \Gamma]\|_F^2 = \sum_{i,j} (\lambda_i - \lambda_j)^2 \left(V^\top A V \right)_{ij}^2. \tag{32}$$

Rotations between directions with different second-order mass incur larger penalties. This is a local compatibility constraint: a zero commutator with the summed generator does not generally imply invariance under a finite product of overlapping-plane rotations.

E Prediction and Online Optimization

E.1 Temporal Heads and Adaptive Fusion

A bias-free cosine classifier has parameters $\phi = (V, \gamma)$ and logits $\ell_{\phi,c}(z) = s\widehat{z}^{\top}\widehat{v}_c$, where hats denote ℓ_2 normalization and $s = \mathrm{clip}(e^{\gamma}, 1, 100)$. After optimizer step k, only the active interval's EMA heads are updated:

$$\phi_{e,m}^{(k)} = \alpha_m \phi_{e,m}^{(k-1)} + (1-\alpha_m)\phi_{\mathrm{pl}}^{(k)}. \tag{33}$$

Both class weights and the log-scale are averaged, without backpropagation into EMA parameters. Each state is an exponentially weighted history of plastic parameters. The decay rates 0.9 and 0.99 correspond to characteristic horizons of 10 and 100 optimizer steps, respectively. With E active intervals, prediction uses EM EMA heads and one plastic head on the same adapted feature z.

The plastic and EMA predictions are $p_{\mathrm{pl}} = \mathrm{SoftMax}(\ell_{\mathrm{pl}})$ and $p_{e,m} = \mathrm{SoftMax}(\ell_{e,m})$. Stable responses are constructed as

$$\begin{aligned} s_{e,c} &= \max_{m=1,\dots,M} p_{e,m,c}, \\ s_{\mathrm{st}} &= \sum_{e=1}^{E} r_e s_e. \end{aligned} \tag{34}$$

Historical aggregation always uses r, including during training when prompts use $w^{\mathrm{tr}} = (1-\beta)\mathbf{e}_e + \beta r$, where $\mathbf{e}_e$ denotes the active interval's one-hot vector. The fusion rule is

$$\begin{aligned} g_\nu &= \sigma\left(\frac{\nu - \tau_\nu}{T_\nu}\right), \\ \Delta &= \left[\max_c p_{\mathrm{pl},c} - \max_c s_{\mathrm{st},c}\right]_+, \\ g &= 1 - (1-g_\nu)e^{-\kappa\Delta}, \\ s_{\mathrm{final}} &= g p_{\mathrm{pl}} + (1-g)s_{\mathrm{st}}. \end{aligned} \tag{35}$$

All responses and confidence comparisons are computed before external class masking. Inference selects the largest final response among observed classes.

Proposition E.1 (Gate validity and response normalization)**.** *For $T_\nu, \kappa > 0$ and $\Delta \geq 0$, the gate satisfies $g_\nu \leq g \leq 1$ and is nondecreasing in either scalar input with the other held fixed. For normalized head predictions and simplex associations, the stable response mass satisfies*

$$1 \leq Z_{\mathrm{st}} := \sum_c s_{\mathrm{st},c} \leq \min(M, C), \tag{36}$$

where C is the number of output classes. The normalized final distribution can therefore be written as

$$\begin{aligned} \frac{s_{\mathrm{final}}}{\sum_c s_{\mathrm{final},c}} &= \omega p_{\mathrm{pl}} + (1-\omega)\frac{s_{\mathrm{st}}}{Z_{\mathrm{st}}}, \\ \omega &= \frac{g}{g + (1-g)Z_{\mathrm{st}}} \leq g. \end{aligned} \tag{37}$$

Derivation. The factor $(1-g_\nu)e^{-\kappa\Delta}$ lies in $[0, 1-g_\nu]$, establishing the gate bounds. Its partial derivatives are

$$\begin{aligned} \frac{\partial g}{\partial \nu} &= \frac{e^{-\kappa\Delta} g_\nu (1-g_\nu)}{T_\nu} \geq 0, \\ \frac{\partial g}{\partial \Delta} &= \kappa(1-g_\nu)e^{-\kappa\Delta} \geq 0. \end{aligned} \tag{38}$$

For each interval, the class-wise maximum has mass at least that of any individual head and at most the sum of all head masses. Specifically, for any fixed m_0,

$$\begin{aligned} \sum_c \max_m p_{e,m,c} &\geq \sum_c p_{e,m_0,c} = 1, \\ \sum_c \max_m p_{e,m,c} &\leq \sum_c \sum_{m=1}^{M} p_{e,m,c} = M. \end{aligned} \tag{39}$$

Each response is also at most one, giving the upper bound C. Convex averaging over intervals preserves these bounds. The final response has total mass $g + (1-g)Z_{\mathrm{st}}$; division by this quantity yields equation 37. Since $Z_{\mathrm{st}} \geq 1$, the normalized plastic coefficient satisfies $\omega \leq g$.

Thus, g weights response vectors rather than two already normalized distributions. For numerical stability, prediction uses $\log \widetilde{s}_{\mathrm{final}}$, with $\widetilde{s}_{\mathrm{final},c} = \max(s_{\mathrm{final},c}, 10^{-8})$. For a retained class set $\mathcal{C}_{\mathrm{ret}}$ containing the target y, the fused cross-entropy is

$$\mathrm{CE}\left(\log \widetilde{s}_{\mathrm{final}} + m, y\right) = -\log \frac{\widetilde{s}_{\mathrm{final},y}}{\sum_{c \in \mathcal{C}_{\mathrm{ret}}} \widetilde{s}_{\mathrm{final},c}}. \tag{40}$$

The response-mass bound concerns the ideal responses before numerical flooring. These properties establish the validity of the fusion computation without implying calibration or accuracy dominance over both constituent branches.

E.2 Prototype Memory and the Complete Loss

For class c, the memory stores a count n_c and the cumulative mean μ_c of normalized adapted features. After each optimizer step, the preceding forward pass supplies detached normalized features $\widehat{z}_i$ for the current class subset $\mathcal{B}_c$. Starting from $\mu_c = 0$ and $n_c = 0$, the update is

$$\begin{aligned} \mu_c &\leftarrow \frac{n_c \mu_c + \sum_{i \in \mathcal{B}_c} \widehat{z}_i}{n_c + |\mathcal{B}_c|}, \\ n_c &\leftarrow n_c + |\mathcal{B}_c|. \end{aligned} \tag{41}$$

Classes absent from the batch retain their previous states. The update is a running mean by induction, with $\|\mu_c\|_2 \leq 1$. Counts refer to feature observations across optimization steps, not unique images.

For the available class set $\mathcal{C}_{\mathrm{mem}} = \{c : n_c > 0\}$, prototype supervision is evaluated when this set is nonempty:

$$\mathcal{L}_{\mathrm{proto}} = \frac{1}{|\mathcal{C}_{\mathrm{mem}}|} \sum_{c \in \mathcal{C}_{\mathrm{mem}}} \mathrm{CE}\left(\ell_{\mathrm{pl}}\big(\mathrm{sg}(\mu_c)\big), c\right). \tag{42}$$

Each available class contributes one supervised prototype, without an output mask; the softmax includes all C classifier outputs. This loss directly updates only the plastic classifier, rather than attracting image features to class centers. Prototypes are used for training only.

For the image batch, the mask m retains its present classes, or all observed classes when batch masking is disabled. The sample-based objective is

$$\mathcal{L}_{\mathrm{data}} = 0.7\, \mathbb{E}_{(x,y)\in\mathcal{B}}\, \mathrm{CE}\left(\ell_{\mathrm{pl}}(x) + m, y\right) + 0.3\, \mathbb{E}_{(x,y)\in\mathcal{B}}\, \mathrm{CE}\left(\log \widetilde{s}_{\mathrm{final}}(x) + m, y\right). \tag{43}$$

Combining sample and prototype supervision gives

$$\mathcal{L}_{\mathrm{cls}} = \begin{cases} \frac{1}{2}\left(\mathcal{L}_{\mathrm{data}} + \mathcal{L}_{\mathrm{proto}}\right), & \mathcal{C}_{\mathrm{mem}} \neq \varnothing, \\ \mathcal{L}_{\mathrm{data}}, & \mathcal{C}_{\mathrm{mem}} = \varnothing. \end{cases} \tag{44}$$

The complete online objective is

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{con}} \bar{f}_{\mathcal{B}} \mathcal{L}_{\mathrm{con}} + \lambda_{\mathrm{geo}} \mathcal{L}_{\mathrm{geo}}, \\ \bar{f}_{\mathcal{B}} &= \mathrm{sg}\left(\frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \big(1 - \nu(x)\big)\right). \end{aligned} \tag{45}$$

With memory available, plastic, fused, and prototype losses have coefficients 0.35, 0.15, and 0.50, respectively. The loss uses existing prototypes; after optimization, the active EMA heads and then prototypes are updated. After all K local steps, fixed-feature statistics are collected once, followed by the interval check.

Image-based losses update shared prompts, lifting maps, active angles, and the plastic head. Fixed EMA heads still transmit gradients to their input features. Measurement responses, prototypes, and

Algorithm 1 MIST: online adaptation and continual prediction

Require: Stream $\mathcal{D}$, frozen backbone f_Θ, fixed projection R, Givens planes, sample boundaries, EMA rates α_m, local steps K

Initialize shared Z_0^l, W^l, plastic state ϕ_{pl}, $(\mu_c, n_c) = (0, 0)$, and $\mathcal{C}_{\mathrm{seen}} = \varnothing$
Activate $E = e = 1$: $(H_e, N_e) = (0, 0)$, $\theta_e^l = 0$, and $\phi_{e,m} = \phi_{\mathrm{pl}}$
(1) Online training
for each incoming mini-batch (X, y) **do**
  Register classes, remap labels, and construct the retained-class mask m
  for $k = 1, \ldots, K$ **do**
    $\widetilde{X} \leftarrow \mathrm{Augment}(X)$; $(z, \ell, s, \nu) \leftarrow \mathrm{Forward}(\widetilde{X}, \mathrm{train})$
    Recompute detached Γ_j^l for $j < e$; evaluate the penalties in equation 27
    Evaluate prototype supervision from existing memory and optimize equation 45
    $\phi_{e,m} \leftarrow \alpha_m \phi_{e,m} + (1 - \alpha_m)\phi_{\mathrm{pl}}$ for all m
    Update (μ_c, n_c) using cached detached features by equation 41
  end for
  $Q \leftarrow \mathrm{Normalize}\big(f_\Theta^{\mathrm{cls}}(\mathrm{Transform}_{\mathrm{eval}}(X))R\big)$
  $H_e \leftarrow H_e + Q^\top Q$; $N_e \leftarrow N_e + |X|$; advance the sample count
  if the scheduler activates the next interval **then**
    $e \leftarrow e + 1$; $E \leftarrow e$; $(H_e, N_e) \leftarrow (0, 0)$
    $\theta_e^l \leftarrow (e-1)^{-1} \sum_{j<e} \theta_j^l + 0.005\xi^l$, with $\xi^l \sim \mathcal{N}(0, I)$
    $\phi_{e,m} \leftarrow \phi_{\mathrm{pl}}$ for all m
  end if
end for
(2) Inference
$(z, \ell, s, \nu) \leftarrow \mathrm{Forward}(x, \mathrm{test})$; $\hat{y} \leftarrow \arg\max_{c \in \mathcal{C}_{\mathrm{seen}}} s_c$
Shared computation: $\mathrm{Forward}(x, \chi)$
$q \leftarrow \mathrm{Normalize}\big(f_\Theta^{\mathrm{cls}}(x)R\big)$; $C_j \leftarrow H_j / \max(N_j, 1)$
$\rho_j \leftarrow (C_j + \epsilon I)/\operatorname{Tr}(C_j + \epsilon I)$; $\rho_0 \leftarrow I/d_m$
$\pi_j \leftarrow (1 - \pi_0)(N_j + 1)/\sum_{a=1}^{E}(N_a + 1)$, for $j = 1, \ldots, E$
$S \leftarrow \sum_{j=0}^{E} \pi_j \rho_j$; $a_j \leftarrow q S^{-1/2}(\pi_j \rho_j) S^{-1/2} q^\top$, for $j = 0, \ldots, E$
$\nu \leftarrow a_0$; $r_j \leftarrow a_j / \sum_{a=1}^{E} a_a$; $w \leftarrow (1 - \beta)\mathbf{e}_e + \beta r$ if training, otherwise $w \leftarrow r$
$P^l \leftarrow Z_0^l \left[\prod_r G_r \left(\sum_j w_j \theta_{j,r}^l\right)\right] W^l + B$
$z \leftarrow f_\Theta(x; \{P^l\})$; $\ell \leftarrow \ell_{\phi_{\mathrm{pl}}}(z)$; $p \leftarrow \mathrm{SoftMax}(\ell)$
$s_{\mathrm{st}} \leftarrow \sum_j r_j \max_m \mathrm{SoftMax}\big(\ell_{\phi_{j,m}}(z)\big)$
$\Delta \leftarrow [\max_c p_c - \max_c s_{\mathrm{st},c}]_+$; $g_\nu \leftarrow \sigma((\nu - \tau_\nu)/T_\nu)$
$g \leftarrow 1 - (1 - g_\nu)e^{-\kappa\Delta}$; $s \leftarrow gp + (1 - g)s_{\mathrm{st}}$; return (z, ℓ, s, ν)

reference moments are detached; geometric penalties directly affect active angles. Historical angle slices receive no direct loss gradient. Because they share a parameter tensor with active angles, strict immutability additionally requires handling inactive-slice optimizer momentum and weight decay.

Prototype memory summarizes features collected under evolving prompt states, whereas measurement memory uses the fixed, unprompted representation. The prototypes therefore provide approximate historical summaries without storing past raw images.

F Online Procedure and Complexity

Algorithm 1 uses a shared forward computation for training and inference. The measurement computation applies the numerical safeguards described in Appendix D.1. The index e denotes the active internal interval, and E denotes the number of instantiated intervals.

Complexity. The following costs describe the dense implementation and exclude backbone computation, optimizer states, and temporary activations. Here, $L = |\mathcal{L}|$, B_{sz} is the batch size, and C is the number of output classes.

- **Persistent memory.** The fixed projection, interval statistics, shared prompts, angle banks, classifiers, and prototypes require

$$O\big(Dd_m + E_{\max}d_m^2 + Ld_p(P+D) + E_{\max}LR_g + (E_{\max}M+2)CD\big). \tag{46}$$

  Banks are allocated for the maximum interval capacity $E_{\max}$.
- **Measurement.** Mixture construction and factorization cost $O(Ed_m^2 + d_m^3)$ per forward batch. Projection and quadratic responses cost $O(Dd_m + Ed_m^2)$ per sample. Statistical collection costs $O(B_{\text{sz}}(Dd_m + d_m^2))$ per incoming batch, excluding backbone feature extraction.
- **Prompt construction and prediction.** The per-sample cost is

$$O\big(L\big(E_{\max}R_g + R_g d_p^3 + Pd_p^2 + Pd_pD\big) + (EM+1)CD\big). \tag{47}$$

  The cubic rotation term arises from composing dense Givens matrices.
- **Training-only additions.** Prototype supervision costs $O(|\mathcal{C}_{\text{mem}}|CD)$, and active-bank EMA updates cost $O(MCD)$. Historical moment construction and commutator evaluation add

$$O\left((E-1)L\left(E_{\max}R_g + (R_g+1)d_p^3 + Pd_p^2\right)\right) \tag{48}$$

  per optimizer step.

Inference uses one unprompted and one prompted backbone pass. Each incoming training batch additionally requires an unprompted collection pass after its K adaptation steps. No final offline solve is required.

G Additional Experiment Results

G.1 Comparison with Offline CL Methods.

To further assess MIST beyond methods specifically developed for GCL, we compare it with eight conventional CL methods: S-Prompt++ (Wang et al., 2022b), HiDe (Wang et al., 2023), NoRGa (Le et al., 2024), SD-LoRA (Wu et al., 2025), COSO (Cheng et al., 2025), and DualLoRA (Chen et al., 2026). As shown in Table 5, MIST achieves the highest mean A_{auc} and A_{last} on CIFAR-100 and CUB-200. On CIFAR-100, MIST obtains 82.98% A_{auc} and 86.61% A_{last}, compared with 80.21% and 85.67% for the strongest respective baselines. The advantage is more pronounced on CUB-200, where MIST achieves 76.98% A_{auc} and 80.58% A_{last}, exceeding the strongest respective baselines at 67.26% and 71.01%. On ImageNet-R, MIST achieves 51.96% A_{auc} and 50.68% A_{last}, below the highest reported values of 55.51% and 53.55%. Overall, MIST maintains competitive performance against the evaluated conventional CL methods, with a particularly clear advantage on fine-grained recognition. We further compare MIST with CoSo and DualLoRA under the self-supervised iBOT-21K and iBOT-1K backbones, as shown in Fig. 4.

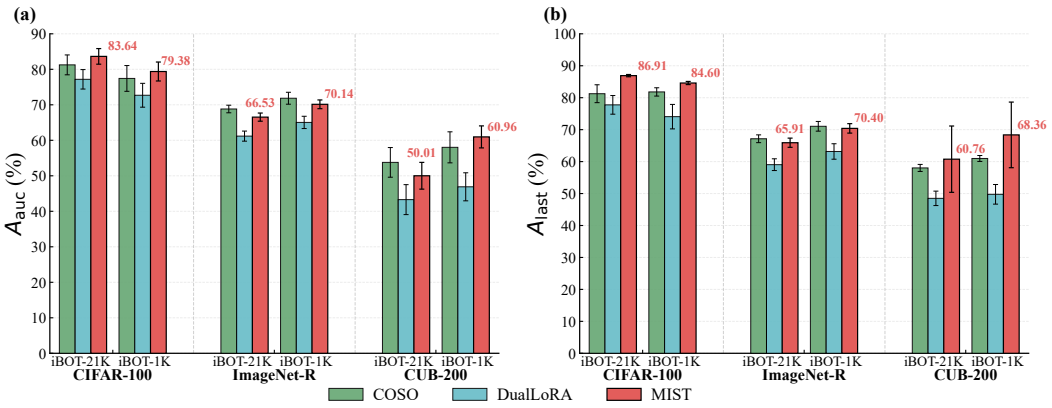

Figure 4: Comparison with offline CL methods under self-supervised iBOT-21K and iBOT-1K backbones. (a) A_{auc} and (b) A_{last} on CIFAR-100, ImageNet-R, and CUB-200. Error bars denote the standard deviation over five runs.

Table 5: Comparison with prominent offline methods on three GCL benchmarks under Sup-21K.

Method	CIFAR-100		ImageNet-R		CUB-200	
	$A_{\text{auc}} \uparrow$	$A_{\text{last}} \uparrow$	$A_{\text{auc}} \uparrow$	$A_{\text{last}} \uparrow$	$A_{\text{auc}} \uparrow$	$A_{\text{last}} \uparrow$
S-Prompt++	80.21 ± 2.55	83.48 ± 1.20	52.14 ± 1.65	49.13 ± 1.60	66.61 ± 2.21	64.73 ± 2.25
HiDe-Prompt	77.10 ± 3.81	81.77 ± 2.00	53.77 ± 1.09	49.87 ± 3.01	67.05 ± 2.37	67.12 ± 0.50
HiDe-LoRA	80.07 ± 2.41	82.00 ± 1.25	55.09 ± 1.45	51.29 ± 6.29	67.26 ± 1.76	67.28 ± 1.45
HiDe-Adapter	79.52 ± 2.81	81.41 ± 0.95	53.92 ± 1.32	50.86 ± 5.08	66.09 ± 1.41	64.53 ± 1.78
NoRGa	78.89 ± 3.33	83.03 ± 1.20	54.12 ± 1.37	50.09 ± 3.66	67.16 ± 2.44	67.06 ± 0.58
SD-LoRA	79.26 ± 2.21	78.91 ± 2.48	$\mathbf{55.51} \pm 1.30$	51.97 ± 3.09	64.12 ± 2.02	60.57 ± 0.77
COSO	64.17 ± 0.61	79.84 ± 0.96	39.85 ± 1.02	40.83 ± 0.66	48.28 ± 1.10	57.36 ± 2.05
DualLoRA	78.07 ± 0.62	85.67 ± 0.43	50.54 ± 1.40	$\mathbf{53.55} \pm 0.68$	58.51 ± 1.13	71.01 ± 1.32
MIST	$\mathbf{82.98} \pm 1.59$	$\mathbf{86.61} \pm 0.35$	51.96 ± 1.48	50.68 ± 1.00	$\mathbf{76.98} \pm 2.02$	$\mathbf{80.58} \pm 1.54$

G.2 Effect of Task Granularity in Si-Blurry

Tab. 6, Fig. 5, and Tab. 7 examine the sensitivity of GCL methods to task granularity under both supervised Sup-21K and self-supervised iBOT-21K pretraining.

Under Sup-21K, MIST consistently achieves the highest mean A_{auc} and A_{last} on CIFAR-100 and CUB-200 across all task counts. On CIFAR-100, its A_{auc} decreases moderately from 82.98% to 79.19% as the number of tasks increases from 5 to 20, while A_{last} remains above 84%. The advantage is more pronounced on CUB-200. Compared with FlyPrompt, the A_{auc} margin increases from 6.34 points with 5 tasks to 14.94 points with 20 tasks, while the A_{last} margin increases from 7.18 to 18.40 points. ImageNet-R exhibits a different trend: MIST is below FlyPrompt under 5 tasks, but achieves the highest mean A_{last} for 10, 15, and 20 tasks and the highest A_{auc} under 20 tasks. These results show that the relative benefit of MIST becomes more evident under finer stream partitions, particularly for fine-grained recognition.

The self-supervised iBOT-21K results further strengthen this observation. MIST achieves the highest mean A_{auc} and A_{last} for every evaluated task number on all three benchmarks. On CIFAR-100, MIST improves over FlyPrompt by 8.06–11.33 points in A_{auc} and 7.55–9.97 points in A_{last} across the four task settings. On ImageNet-R, the corresponding gains reach 8.78–17.44 and 11.52–21.52 points, respectively. The largest differences occur on CUB-200: with 10 tasks, MIST reaches 48.30% A_{auc} and 62.00% A_{last}, compared with 20.03% and 21.69% for FlyPrompt. Even under 20 tasks, where the fine-grained stream becomes substantially more challenging, MIST retains 28.37% A_{auc} and 39.91% A_{last}, remaining clearly above the competing methods.

Overall, increasing task granularity generally makes continual adaptation more difficult, especially under self-supervised pretraining and on CUB-200. Nevertheless, MIST maintains a consistent advantage across different stream partitions and pretraining paradigms. These results support the effectiveness of retaining reusable adaptation knowledge in a shared latent prompt while expressing distribution-dependent variation through measurement-conditioned isometric transport, rather than relying on increasingly fragmented interval-specific prompt states.

G.3 Effect of the Disjoint Class Ratio

Tab.8 evaluates different disjoint class ratios with $r_B = 10\%$ and the Sup-21K-pretrained backbone. The three configurations cover fully blurry, mixed, and fully disjoint class exposure, testing adaptation under different degrees of class recurrence.

MIST achieves the highest mean A_{auc} on CIFAR-100 and the highest mean A_{auc} and A_{last} on CUB-200 across all evaluated ratios. Its final accuracy also remains relatively stable on these two benchmarks as the proportion of disjoint classes changes. On CUB-200, the advantage over the strongest competing method is largest in the mixed setting and remains present under fully disjoint exposure. These observations support the applicability of shared adaptation to streams containing both recurring and new classes, without requiring cross-session class recurrence for the framework to remain effective.

Table 6: Performance comparison of GCL methods under different numbers of tasks on CIFAR-100, ImageNet-R, and CUB-200 using the Sup-21K-pretrained ViT-B/16 backbone. Results are reported as mean ± standard deviation over five random seeds (%).

Method	5 Tasks		10 Tasks		15 Tasks		20 Tasks	
	A_{auc}	A_{last}	A_{auc}	A_{last}	A_{auc}	A_{last}	A_{auc}	A_{last}
CIFAR-100								
L2P	76.23 ± 2.73	79.11 ± 1.43	70.79 ± 3.23	77.79 ± 1.02	69.61 ± 2.26	74.67 ± 0.47	68.15 ± 1.11	74.07 ± 0.45
DualPrompt	76.04 ± 3.32	76.62 ± 0.74	74.77 ± 2.72	81.16 ± 0.75	74.74 ± 2.14	79.11 ± 0.45	73.47 ± 2.52	77.23 ± 0.02
CODA-P	79.13 ± 3.06	80.91 ± 0.70	68.59 ± 6.76	74.18 ± 5.15	67.22 ± 7.85	70.92 ± 6.02	63.35 ± 5.98	67.51 ± 5.26
MVP	67.74 ± 4.96	63.22 ± 0.69	58.23 ± 3.42	61.13 ± 6.00	57.16 ± 2.27	55.87 ± 2.41	56.52 ± 3.20	56.87 ± 5.18
MISA	80.35 ± 2.39	80.75 ± 1.24	75.71 ± 3.10	80.22 ± 0.47	77.37 ± 1.65	80.82 ± 0.49	73.96 ± 0.72	77.98 ± 1.17
FlyPrompt	82.65 ± 2.56	85.96 ± 0.70	77.65 ± 3.05	84.87 ± 0.50	77.91 ± 1.98	83.27 ± 0.96	75.87 ± 1.93	81.98 ± 1.12
MIST	**82.98** ± 1.59	**86.61** ± 0.35	**81.06** ± 1.39	**85.52** ± 0.27	**80.45** ± 1.41	**84.77** ± 0.55	**79.19** ± 0.94	**84.05** ± 0.53
ImageNet-R								
L2P	44.40 ± 1.03	42.03 ± 1.72	42.81 ± 1.72	39.39 ± 0.65	39.05 ± 1.66	36.71 ± 0.98	36.13 ± 1.27	35.18 ± 1.47
DualPrompt	46.13 ± 1.94	40.80 ± 1.04	49.35 ± 1.77	45.79 ± 0.82	45.23 ± 1.80	43.30 ± 0.97	42.79 ± 1.45	42.21 ± 0.98
CODA-P	51.87 ± 2.81	48.09 ± 2.75	45.52 ± 2.81	39.84 ± 3.10	37.66 ± 4.54	33.48 ± 4.03	34.36 ± 2.04	31.11 ± 1.47
MVP	39.50 ± 1.41	32.63 ± 3.95	32.84 ± 2.11	26.37 ± 2.74	28.10 ± 2.14	23.24 ± 1.36	27.59 ± 1.39	23.34 ± 2.10
MISA	51.52 ± 2.09	45.08 ± 1.43	**54.59** ± 1.69	49.83 ± 0.63	**50.14** ± 1.60	46.36 ± 0.65	48.01 ± 1.83	45.59 ± 0.96
FlyPrompt	**56.58** ± 1.47	**55.27** ± 0.91	53.98 ± 1.87	50.76 ± 0.97	49.18 ± 1.44	47.41 ± 0.67	47.36 ± 1.35	45.81 ± 1.01
MIST	51.96 ± 1.48	50.68 ± 1.00	53.40 ± 1.72	**51.07** ± 0.83	49.80 ± 1.48	**49.02** ± 1.34	**48.32** ± 1.91	**48.40** ± 0.97
CUB-200								
L2P	64.30 ± 2.18	61.42 ± 2.13	59.44 ± 1.30	57.45 ± 2.26	57.10 ± 3.75	53.95 ± 1.33	53.58 ± 2.38	52.37 ± 2.19
DualPrompt	65.03 ± 2.24	62.43 ± 1.78	61.18 ± 1.39	58.89 ± 2.12	59.16 ± 3.84	56.11 ± 1.51	55.57 ± 2.26	54.14 ± 2.57
CODA-P	66.01 ± 2.20	62.90 ± 2.46	60.68 ± 2.14	58.19 ± 2.00	57.54 ± 4.55	53.82 ± 1.95	53.11 ± 2.51	51.66 ± 2.34
MVP	54.69 ± 3.14	50.07 ± 3.86	46.15 ± 0.62	43.07 ± 4.16	45.28 ± 2.09	39.80 ± 1.00	43.55 ± 2.16	40.89 ± 3.10
MISA	65.40 ± 3.01	60.20 ± 1.82	63.26 ± 1.61	60.58 ± 2.29	60.90 ± 3.79	57.00 ± 0.95	57.71 ± 2.63	55.17 ± 2.52
FlyPrompt	70.64 ± 2.85	73.40 ± 1.88	66.04 ± 2.21	68.11 ± 0.69	62.48 ± 4.19	63.13 ± 1.97	57.76 ± 2.48	57.75 ± 1.00
MIST	**76.98** ± 2.02	**80.58** ± 1.54	**75.52** ± 1.51	**78.72** ± 1.28	**73.90** ± 2.21	**76.99** ± 0.93	**72.70** ± 1.20	**76.15** ± 1.03

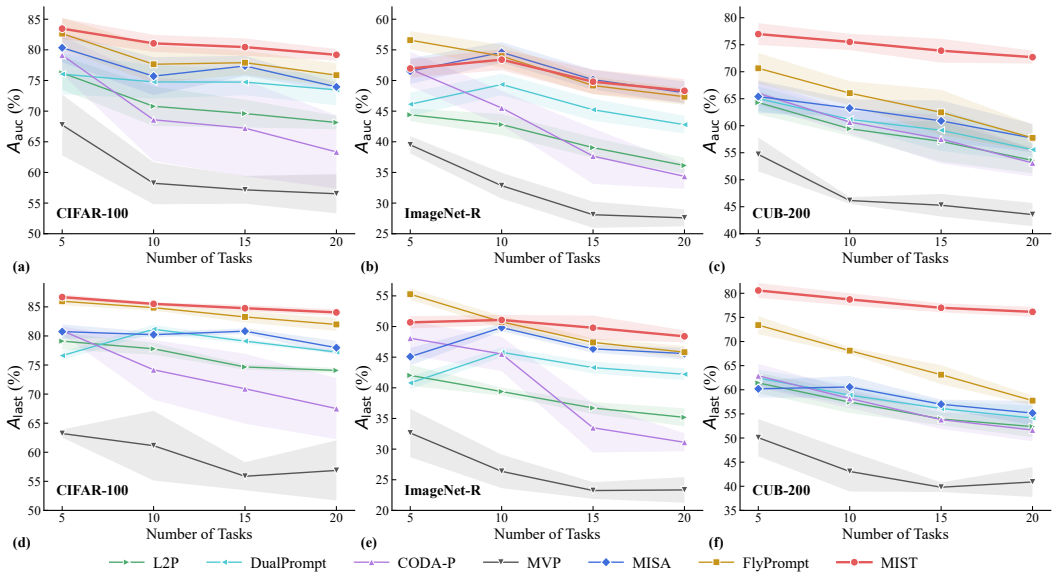

Figure 5: GCL performance with varying task numbers under Sup-21K. (a)–(c) show A_{auc} and (d)–(f) show A_{last} on CIFAR-100, ImageNet-R, and CUB-200, respectively. Lines and shaded regions indicate the mean and standard deviation over five seeds.

The results on ImageNet-R are less favorable: FlyPrompt leads both metrics at every ratio, while MIST consistently ranks second in final accuracy. Thus, changing the degree of class overlap does not eliminate the performance gap on this benchmark. Across methods, the fully disjoint setting generally yields higher A_{auc} without a corresponding improvement in A_{last}. This distinction highlights that online accuracy also depends on the class-exposure structure of the stream; an increase across different r_D settings should therefore not be interpreted alone as improved knowledge retention.

Table 7: Performance comparison of GCL methods under different numbers of tasks on CIFAR-100, ImageNet-R, and CUB-200 using the iBOT-21K-pretrained ViT-B/16 backbone. Results are reported as mean ± standard deviation over five random seeds (%).

Method	5 Tasks		10 Tasks		15 Tasks		20 Tasks	
	A_{auc}	A_{last}	A_{auc}	A_{last}	A_{auc}	A_{last}	A_{auc}	A_{last}
CIFAR-100								
L2P	56.82 ± 8.42	67.61 ± 8.76	51.33 ± 6.56	64.01 ± 4.17	48.64 ± 3.20	60.60 ± 2.39	46.41 ± 1.60	57.42 ± 2.85
DualPrompt	66.06 ± 4.55	67.14 ± 8.60	63.59 ± 5.48	67.84 ± 4.84	63.29 ± 1.80	67.41 ± 2.50	61.35 ± 2.65	63.84 ± 4.80
CODA-P	62.13 ± 7.17	63.38 ± 7.98	58.79 ± 6.98	63.95 ± 5.00	52.65 ± 11.63	58.15 ± 9.34	52.63 ± 7.87	54.81 ± 7.10
MVP	62.33 ± 3.06	48.32 ± 11.42	51.85 ± 6.83	44.87 ± 16.96	50.17 ± 2.24	33.32 ± 4.52	48.12 ± 4.13	35.78 ± 5.11
MISA	65.30 ± 2.28	67.43 ± 6.75	62.94 ± 6.11	68.52 ± 4.97	61.10 ± 2.43	65.52 ± 3.22	60.84 ± 1.80	65.32 ± 2.47
FlyPrompt	75.58 ± 1.70	79.36 ± 3.47	71.26 ± 4.56	78.27 ± 1.82	68.38 ± 4.11	74.07 ± 2.51	69.37 ± 1.51	72.70 ± 1.05
MIST	**83.64** ± 2.20	**86.91** ± 0.35	**82.08** ± 2.42	**85.82** ± 0.46	**79.71** ± 2.34	**83.93** ± 0.38	**78.48** ± 1.98	**82.67** ± 0.45
ImageNet-R								
L2P	35.97 ± 1.62	36.95 ± 2.44	34.59 ± 0.94	32.85 ± 1.71	30.48 ± 1.91	32.34 ± 2.21	26.01 ± 1.48	29.30 ± 1.28
DualPrompt	42.48 ± 1.62	35.91 ± 0.88	43.26 ± 1.80	35.00 ± 3.04	38.93 ± 0.88	32.15 ± 1.61	37.43 ± 1.19	32.94 ± 1.61
CODA-P	45.50 ± 1.66	39.44 ± 1.35	41.93 ± 1.41	34.52 ± 2.32	36.09 ± 3.30	31.00 ± 2.94	34.58 ± 3.88	29.93 ± 2.91
MVP	41.55 ± 1.98	29.29 ± 5.03	33.11 ± 1.86	20.42 ± 6.76	29.56 ± 1.24	19.77 ± 5.03	27.97 ± 1.02	19.08 ± 3.44
MISA	40.94 ± 1.22	36.16 ± 1.58	42.09 ± 1.27	35.59 ± 1.93	36.87 ± 2.38	32.50 ± 2.57	35.03 ± 1.50	32.34 ± 1.23
FlyPrompt	57.75 ± 2.12	54.39 ± 1.29	51.65 ± 1.21	45.48 ± 1.86	46.65 ± 1.72	41.37 ± 2.74	44.60 ± 2.64	40.77 ± 1.05
MIST	**66.53** ± 1.15	**65.91** ± 1.44	**67.02** ± 1.67	**63.95** ± 1.04	**64.09** ± 2.22	**62.89** ± 0.93	**61.64** ± 2.29	**60.80** ± 0.24
CUB-200								
L2P	14.76 ± 1.53	24.51 ± 4.82	10.87 ± 0.73	20.43 ± 2.34	9.39 ± 1.64	15.06 ± 1.99	8.05 ± 1.64	11.80 ± 1.30
DualPrompt	19.90 ± 3.68	21.84 ± 2.35	18.33 ± 1.11	23.85 ± 3.13	14.56 ± 1.28	16.17 ± 3.40	11.36 ± 1.78	10.77 ± 2.62
CODA-P	17.72 ± 5.33	20.82 ± 7.66	15.45 ± 2.45	16.31 ± 3.21	12.63 ± 1.60	9.85 ± 2.27	11.12 ± 1.66	9.52 ± 1.61
MVP	28.73 ± 3.13	23.62 ± 9.51	18.85 ± 2.13	14.42 ± 4.54	15.57 ± 0.99	7.33 ± 3.37	10.68 ± 2.21	3.86 ± 4.08
MISA	18.62 ± 3.36	23.66 ± 2.01	16.77 ± 1.27	20.23 ± 3.06	13.70 ± 2.18	13.13 ± 2.39	10.77 ± 1.20	9.79 ± 2.43
FlyPrompt	28.86 ± 5.84	36.79 ± 7.58	20.03 ± 1.22	21.69 ± 0.62	15.11 ± 1.99	13.36 ± 1.75	13.11 ± 1.77	11.92 ± 1.66
MIST	**50.01** ± 3.77	**60.76** ± 10.38	**48.30** ± 2.31	**62.00** ± 1.44	**40.27** ± 3.94	**55.01** ± 2.01	**28.37** ± 11.30	**39.91** ± 18.47

Table 8: Performance comparison under different r_D settings with $r_B = 10\%$ using the Sup-21K-pretrained backbone. Results are reported as mean ± standard deviation over five random seeds (%). The highest mean for each metric under each setting is highlighted in bold.

r_D (%)	Method	CIFAR-100		ImageNet-R		CUB-200	
		$A_{\text{auc}} \uparrow$	$A_{\text{last}} \uparrow$	$A_{\text{auc}} \uparrow$	$A_{\text{last}} \uparrow$	$A_{\text{auc}} \uparrow$	$A_{\text{last}} \uparrow$
0	L2P	73.64 ± 1.69	81.71 ± 1.28	42.97 ± 1.91	42.53 ± 0.79	64.62 ± 3.37	62.72 ± 3.32
	DualPrompt	72.16 ± 2.40	77.52 ± 1.43	45.08 ± 3.48	42.34 ± 1.55	65.42 ± 3.53	62.84 ± 2.21
	CODA-P	72.33 ± 5.84	78.56 ± 4.46	50.31 ± 4.42	48.88 ± 3.46	66.64 ± 3.42	63.70 ± 2.48
	MVP	67.10 ± 2.33	74.94 ± 7.45	39.04 ± 1.49	32.28 ± 3.12	56.86 ± 3.71	55.50 ± 6.71
	MISA	78.38 ± 1.49	83.04 ± 1.44	51.78 ± 3.22	47.64 ± 0.64	67.38 ± 3.25	64.49 ± 2.55
	FlyPrompt	80.12 ± 1.38	$\mathbf{87.11} \pm 0.52$	$\mathbf{55.44} \pm 1.82$	$\mathbf{55.89} \pm 0.86$	71.60 ± 3.49	74.72 ± 1.69
	MIST	$\mathbf{81.40} \pm 0.58$	87.04 ± 0.40	50.12 ± 1.29	52.67 ± 0.63	$\mathbf{76.36} \pm 3.34$	$\mathbf{80.45} \pm 0.82$
50	L2P	76.23 ± 2.73	79.11 ± 1.43	44.40 ± 1.03	42.03 ± 1.72	64.30 ± 2.18	61.42 ± 2.13
	DualPrompt	76.04 ± 3.32	76.62 ± 0.74	46.13 ± 1.94	40.80 ± 1.04	65.03 ± 2.24	62.43 ± 1.78
	CODA-P	79.13 ± 3.06	80.91 ± 0.70	51.87 ± 2.81	48.09 ± 2.75	66.01 ± 2.20	62.90 ± 2.46
	MVP	67.74 ± 4.96	63.22 ± 0.69	39.50 ± 1.41	32.63 ± 3.95	54.69 ± 3.14	50.07 ± 3.86
	MISA	80.35 ± 2.39	80.75 ± 1.24	51.52 ± 2.09	45.08 ± 1.43	65.40 ± 3.01	60.20 ± 1.82
	FlyPrompt	82.65 ± 2.56	85.96 ± 0.70	$\mathbf{56.58} \pm 1.47$	$\mathbf{55.27} \pm 0.91$	70.64 ± 2.85	73.40 ± 1.88
	MIST	$\mathbf{82.98} \pm 1.59$	$\mathbf{86.61} \pm 0.35$	51.96 ± 1.48	50.68 ± 1.00	$\mathbf{76.98} \pm 2.02$	$\mathbf{80.58} \pm 1.54$
100	L2P	82.98 ± 0.72	78.79 ± 0.95	45.48 ± 0.71	43.12 ± 0.75	71.74 ± 3.58	61.52 ± 1.82
	DualPrompt	81.12 ± 2.10	75.94 ± 0.37	46.79 ± 2.00	41.42 ± 0.63	72.81 ± 3.80	62.46 ± 0.97
	CODA-P	82.68 ± 3.99	77.97 ± 3.19	53.01 ± 3.01	49.31 ± 2.87	73.95 ± 4.10	63.90 ± 0.94
	MVP	74.92 ± 1.10	56.17 ± 2.98	40.43 ± 0.51	28.32 ± 5.55	63.41 ± 2.52	43.80 ± 2.60
	MISA	85.67 ± 1.01	81.04 ± 1.02	53.88 ± 1.99	47.63 ± 0.78	74.27 ± 3.56	62.97 ± 1.44
	FlyPrompt	88.25 ± 0.90	85.51 ± 0.64	$\mathbf{57.41} \pm 0.95$	$\mathbf{55.69} \pm 0.33$	78.14 ± 3.62	74.34 ± 0.74
	MIST	$\mathbf{88.63} \pm 0.53$	$\mathbf{86.51} \pm 0.42$	52.20 ± 1.16	52.41 ± 0.61	$\mathbf{80.55} \pm 4.21$	$\mathbf{79.47} \pm 1.03$

G.4 Effect of the Blurry Sample Ratio

Tab.9 evaluates different blurry class ratios with $r_D = 50\%$ and the Sup-21K-pretrained backbone. Varying r_B changes the degree of blurry class exposure while keeping the disjoint class ratio fixed, allowing us to examine sensitivity to different stream compositions.

MIST achieves the highest mean A_{auc} and A_{last} on CIFAR-100 and CUB-200 across all evaluated ratios. Its performance on both benchmarks varies only moderately as r_B increases, without a consistent decline at higher blurry ratios. In particular, MIST maintains a clear accuracy advantage on CUB-200 throughout the evaluated settings. These results indicate that the framework can retain effective continual adaptation across different degrees of blurry class exposure, rather than relying on a specific class-mixture configuration.

On ImageNet-R, FlyPrompt achieves higher accuracy at every evaluated ratio. Although MIST exhibits relatively small performance variations as r_B changes, its accuracy remains below that of FlyPrompt. The results therefore suggest that MIST is comparatively insensitive to the evaluated blurry ratios on CIFAR-100 and CUB-200, while its performance advantage does not extend to ImageNet-R under the Sup-21K backbone.

Table 9: Performance comparison under different r_B settings with $r_D = 50\%$ using the Sup-21K-pretrained backbone. Results are reported as mean ± standard deviation over five random seeds (%). The highest mean for each metric under each setting is highlighted in bold.

r_B (%)	Method	CIFAR-100		ImageNet-R		CUB-200	
		$A_{\text{auc}} \uparrow$	$A_{\text{last}} \uparrow$	$A_{\text{auc}} \uparrow$	$A_{\text{last}} \uparrow$	$A_{\text{auc}} \uparrow$	$A_{\text{last}} \uparrow$
10	L2P	76.23 ± 2.73	79.11 ± 1.43	44.40 ± 1.03	42.03 ± 1.72	64.30 ± 2.18	61.42 ± 2.13
	DualPrompt	76.04 ± 3.32	76.62 ± 0.74	46.13 ± 1.94	40.80 ± 1.04	65.03 ± 2.24	62.43 ± 1.78
	CODA-P	79.13 ± 3.06	80.91 ± 0.70	51.87 ± 2.81	48.09 ± 2.75	66.01 ± 2.20	62.90 ± 2.46
	MVP	67.74 ± 4.96	63.22 ± 0.69	39.50 ± 1.41	32.63 ± 3.95	54.69 ± 3.14	50.07 ± 3.86
	MISA	80.35 ± 2.39	80.75 ± 1.24	51.52 ± 2.09	45.08 ± 1.43	65.40 ± 3.01	60.20 ± 1.82
	FlyPrompt	82.65 ± 2.56	85.96 ± 0.70	$\mathbf{56.58} \pm 1.47$	$\mathbf{55.27} \pm 0.91$	70.64 ± 2.85	73.40 ± 1.88
	MIST	$\mathbf{82.98} \pm 1.59$	$\mathbf{86.61} \pm 0.35$	51.96 ± 1.48	50.68 ± 1.00	$\mathbf{76.98} \pm 2.02$	$\mathbf{80.58} \pm 1.54$
30	L2P	78.48 ± 0.92	80.13 ± 0.87	43.32 ± 0.79	42.27 ± 1.40	63.67 ± 2.03	63.60 ± 3.09
	DualPrompt	77.76 ± 1.65	77.50 ± 0.49	45.11 ± 1.09	41.01 ± 0.70	64.36 ± 1.98	63.72 ± 2.22
	CODA-P	81.50 ± 1.20	82.65 ± 0.72	50.55 ± 2.24	47.58 ± 2.81	65.89 ± 1.42	64.97 ± 2.83
	MVP	71.01 ± 1.70	65.71 ± 4.60	38.62 ± 1.35	32.43 ± 4.75	54.16 ± 4.56	51.04 ± 6.89
	MISA	82.54 ± 1.08	82.50 ± 0.68	51.69 ± 0.94	47.09 ± 1.16	67.13 ± 1.72	66.53 ± 2.39
	FlyPrompt	84.61 ± 1.25	86.89 ± 0.38	$\mathbf{55.30} \pm 0.86$	$\mathbf{55.43} \pm 1.04$	70.19 ± 2.01	74.25 ± 1.27
	MIST	$\mathbf{84.82} \pm 0.58$	$\mathbf{87.60} \pm 0.48$	52.06 ± 1.39	52.06 ± 1.39	$\mathbf{75.73} \pm 1.72$	$\mathbf{81.21} \pm 0.45$
50	L2P	77.44 ± 2.42	80.31 ± 0.69	44.39 ± 1.72	43.66 ± 1.04	65.42 ± 2.71	64.62 ± 1.63
	DualPrompt	77.44 ± 2.64	77.13 ± 1.08	46.23 ± 1.83	41.99 ± 0.72	66.40 ± 2.72	65.34 ± 3.05
	CODA-P	81.39 ± 2.18	83.10 ± 0.97	53.05 ± 1.60	50.20 ± 1.74	67.88 ± 2.26	65.44 ± 2.32
	MVP	67.97 ± 4.78	58.11 ± 1.26	40.68 ± 1.59	31.87 ± 6.56	57.25 ± 3.76	53.86 ± 3.41
	MISA	81.81 ± 2.29	82.51 ± 0.38	53.27 ± 1.71	48.32 ± 0.84	68.68 ± 2.47	66.84 ± 2.27
	FlyPrompt	83.69 ± 1.81	86.31 ± 0.73	$\mathbf{56.50} \pm 1.87$	$\mathbf{55.59} \pm 0.83$	71.95 ± 1.92	74.03 ± 1.02
	MIST	$\mathbf{84.11} \pm 1.55$	$\mathbf{86.55} \pm 0.53$	51.22 ± 1.79	51.48 ± 1.87	$\mathbf{77.13} \pm 1.25$	$\mathbf{81.37} \pm 0.23$