UnL^2: Unlocking Label Likelihoods in Multimodal Large Language Models for Class-Incremental Learning

Anonymous authors
Paper under double-blind review

Abstract

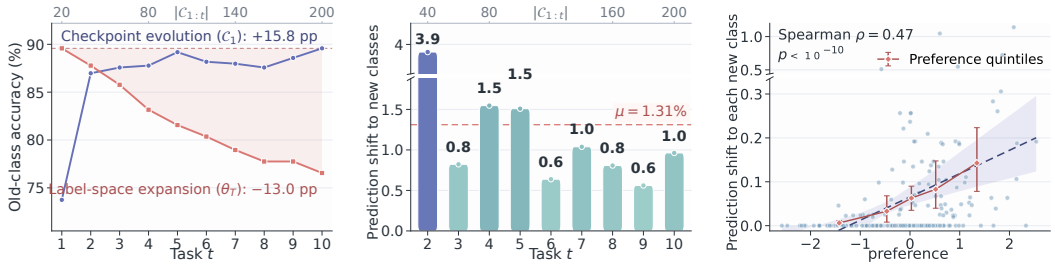
Multimodal large language models (MLLMs) have shown strong potential for class incremental learning (CIL), as flexible semantic generation supports the acquisition of new classes while retaining recognition of previous ones. In this paper, we reveal overlooked inference interference induced by label-space expansion, with persistent generative preference among candidate classes progressively biasing predictions toward newly introduced classes and degrading recognition of previous ones. To address this issue, we propose Unlocking Label Likelihoods (UnL^2), which mitigates inference interference by calibrating preference from generative evidence. Specifically, UnL^2 derives generative evidence over the accumulated label space through Candidate Scoring and dynamically updates a calibration queue using the derived evidence, enabling Online Calibration to estimate a class-wise correction via Sinkhorn matching for subsequent predictions. Theoretical analysis establishes exact absorption of the shared additive class-wise preference component and bounded sensitivity of calibrated pairwise evidence to sample-specific deviations. Extensive experiments on multiple CIL benchmarks show consistent improvements over existing methods, with larger gains as the number of incremental tasks increases.

1 Introduction

Class-Incremental Learning (CIL) aims to progressively acquire new classes from sequential data streams while retaining recognition of previously learned ones (Wang et al., 2023; Zhou et al., 2024b). To this end, early CIL methods mainly train models from scratch and accommodate sequentially arriving classes through rehearsal-based (Shin et al., 2017; Rostami et al., 2019), architecture-based (Yoon et al., 2018), or regularization-based (Liu et al., 2018; Oh et al., 2022) methods. More recently, large-scale pretrained models have become powerful foundations for CIL by transferring rich prior knowledge and generalizable representations to incremental learning (Wang et al., 2022; Liang & Li, 2024; Liu & Chang, 2025; Zhu et al., 2025). In particular, vision-language models (VLMs), such as CLIP (Radford et al., 2021), further exploit cross-modal semantic knowledge for incremental recognition (Liu et al., 2025; Zhang et al., 2026). However, existing pretrained CIL methods remain predominantly discriminative, whereas multimodal large language models (MLLMs) enable a generative alternative through flexible semantic generation (Zhu et al., 2024), supporting new-class acquisition while retaining recognition of previous ones (Cao et al., 2024).

Building on this generative paradigm, GMM (Cao et al., 2024) discards the conventional classifier and performs recognition through free-form generation followed by semantic matching against candidate labels (Hessel et al., 2021). Free-form generation can produce semantically ambiguous responses or responses outside the valid candidate set, particularly for visually or semantically similar classes (Feng et al., 2025; Choi et al., 2026). KG-GMM (Cao et al., 2025) addresses this limitation by introducing an evolving knowledge graph (Ma et al., 2025) to model inter-class relations and enrich generated responses with more discriminative semantics (Chen et al., 2025). Existing generative CIL methods therefore primarily improve recognition by enhancing the semantic quality of generated responses. At each incremental task, model parameters are updated as the accumulated label space expands, making their respective effects on old-class recognition difficult to disentangle.

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.



(a) Performance (b) Prediction Shift Distribution (c) Preference–Shift Association

Figure 1: **Analysis of inference interference induced by label-space expansion on ImageNet-R.** (a) Successive parameter updates evaluated over the same old-class label space consistently improve old-class accuracy, whereas evaluating the final trained model over progressively expanded accumulated label spaces causes substantial degradation. (b) Previously correct old-class predictions shift disproportionately toward newly introduced classes. (c) Candidate classes with stronger generative preference receive a larger fraction of prediction shifts, showing a significant positive correlation between generative preference and prediction redistribution ($\rho = 0.47, p < 10^{-10}$). These phenomena are consistently observed across different MLLM backbones, as shown in Figure 10.

We disentangle these effects through controlled evaluations of parameter updates and label-space expansion. Figure 1(a) shows that successive parameter updates evaluated over the same old-class label space consistently improve old-class accuracy, whereas progressively expanding the accumulated label space under the final trained model causes substantial degradation. These contrasting trends identify label-space expansion as a distinct source of inference interference in generative CIL, separate from the effect of parameter updates. Figure 1(b) further shows that previously correct old-class predictions shift disproportionately toward newly introduced classes rather than being distributed across the accumulated label space. Such non-uniform redistribution suggests systematic class-wise differences in prediction bias, which we characterize by quantifying generative preference for each candidate class and relating it to the shifts received by that class. As shown in Figure 1(c), classes with stronger generative preference receive larger prediction shifts. Together, these results show that persistent generative preference increasingly biases predictions toward newly introduced classes as the accumulated label space expands, sustaining inference interference in generative CIL.

To address this issue, we propose Unlocking Label Likelihoods (UnL^2), which mitigates inference interference by calibrating class-wise generative preference using generative evidence over the accumulated label space. Candidate Scoring replaces free-form generation with length-normalized autoregressive likelihood evaluation over candidate answer sequences, converting unrestricted generation into comparable class-wise evidence while eliminating competition from responses outside the candidate set. This evidence further reveals persistent generative preference among accumulated classes, which increasingly biases predictions toward newly introduced classes as the label space expands. Online Calibration aggregates generative evidence from previously processed samples in a bounded FIFO calibration queue and performs Sinkhorn matching between the predictive class marginal and a target marginal, yielding a class-wise correction for the current sample's generative evidence. Theorem 1 establishes that Sinkhorn calibration exactly absorbs the shared additive class-wise preference component while bounding the effect of sample-specific deviations on calibrated pairwise evidence. Extensive experiments across multiple CIL benchmarks demonstrate consistent improvements over existing methods, with larger gains as the number of incremental tasks increases.

2 Related Work

2.1 Class-Incremental Learning

Early CIL methods primarily train models from scratch, rehearsal-based methods replay stored or generated samples from previous classes during incremental optimization (Shin et al., 2017; Rostami et al., 2019), architecture-based methods allocate isolated model components to different tasks or progressively expand the network (Yoon et al., 2018), while regularization-based methods penalize changes to parameters or predictions that are important for previously learned classes (Liu et al., 2018; Oh et al., 2022). The emergence of large-scale pretrained models (PTMs) has shifted CIL to-

ward exploiting transferable representations with lightweight adaptation. Selection-based methods retrieve task-relevant prompts or pretrained components according to the current input (Wang et al., 2022), prototype-rectification methods correct accumulated class prototypes to compensate for representation drift across incremental tasks (Liu & Chang, 2025), and orthogonality-based methods project gradients onto constrained subspaces to prevent new-task updates from interfering with previously acquired knowledge (Liang & Li, 2024; Zhu et al., 2025). Vision-language models further extend PTM-based CIL by exploiting cross-modal semantic knowledge, with CLIP-based methods adapting transferable image-text representations for recognition over accumulated classes (Radford et al., 2021; Liu et al., 2025; Zhang et al., 2026). More recently, multimodal large language models enable a generative paradigm through flexible semantic generation (Zhu et al., 2024), supporting the acquisition of new classes while retaining recognition of previous ones (Cao et al., 2024; 2025).

2.2 Generative Models for Classification

Recent studies repurpose generative models for visual recognition by deriving class evidence directly from class-conditional generation, thereby avoiding the need for an explicit discriminative classifier. Text-to-image diffusion models instantiate this paradigm by conditioning the denoising process on candidate class descriptions and ranking classes according to class-conditional likelihoods or corresponding noise-prediction errors (Clark & Jaini, 2023; Li et al., 2023a). Subsequent methods further introduce discriminative objectives or energy-based formulations to strengthen the separation between competing classes (Huang et al., 2023; Guo et al., 2023). Multimodal large language models extend this generative classification paradigm to autoregressive language modeling. Open-ended approaches formulate recognition as free-form category verbalization from visual inputs (Conti et al., 2025), while candidate-based approaches evaluate candidate answers through their response similarity (Dai et al., 2023). More recent answer-extraction methods further connect free-form generation with predefined candidate sets by mapping generated responses to candidate labels (Yu et al., 2024b; Lawrence et al., 2026). Despite these advances, existing generative classification methods are predominantly studied under fixed label spaces.

3 Preliminaries

Class-incremental learning. CIL considers a sequence of T tasks with mutually disjoint class increments $\{\Delta\mathcal{C}_t\}_{t=1}^{T}$. At task t, only the current training set $\mathcal{D}_t = \{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{n_t}$ is available, where $\mathbf{x}_i^t$ denotes an image and $y_i^t \in \Delta\mathcal{C}_t$ its corresponding label. The class increments satisfy $\Delta\mathcal{C}_t \cap \Delta\mathcal{C}_{t'} = \varnothing$ for $t \neq t'$, yielding the accumulated label space $\mathcal{C}_t = \bigcup_{j=1}^{t} \Delta\mathcal{C}_j$, with $K_t = |\mathcal{C}_t|$ denoting the number of accumulated classes. Under the exemplar-free setting (Liang & Li, 2024; Huang et al., 2025), previous-task samples are inaccessible during optimization, while task identities are unavailable at inference (Wang et al., 2025), requiring each test sample to be recognized over the entire accumulated label space $\mathcal{C}_t$.

Generative optimization. Following prior generative CIL practice (Cao et al., 2024), recognition is formulated as conditional language generation using a fixed question $\mathbf{q}$, i.e., "What is this photo?". Each class $c \in \mathcal{C}_t$ is represented by a candidate answer sequence $\mathbf{a}^c = (a_1^c, \ldots, a_{L_c}^c)$ of answer length L_c, instantiated with the template "This is a photo of [class].". Given a training sample $\mathbf{x}_i^t$, the model is optimized by minimizing the token-level autoregressive negative log-likelihood

$$\mathcal{L}_{\text{gen}}^t = -\frac{1}{\sum_{i=1}^{n_t} L_{y_i^t}} \sum_{i=1}^{n_t} \sum_{\ell=1}^{L_{y_i^t}} \log p_{\boldsymbol{\theta}^t}\left(a_\ell^{y_i^t} \mid \mathbf{x}_i^t, \mathbf{q}, \mathbf{a}_{<\ell}^{y_i^t}\right), \tag{1}$$

where $\boldsymbol{\theta}^t$ denotes the trainable vision-to-language projection parameters at task t. At each task, $\boldsymbol{\theta}^t$ is initialized from $\boldsymbol{\theta}^{t-1}$ and optimized solely on $\mathcal{D}_t$, while the pretrained visual encoder, query transformer, and language model remain frozen (Cao et al., 2024; 2025).

4 Method

As illustrated in Figure 2, UnL^2 mitigates inference interference induced by label-space expansion by calibrating predictions with generative evidence over the accumulated label space. Candidate

Scoring replaces free-form generation with likelihood evaluation of candidate answer sequences, suppressing preference toward responses outside the accumulated label space while exposing persistent generative preference among candidate classes. Online Calibration tracks such generative preference through a calibration queue and performs Sinkhorn matching between the predictive class marginal and a target marginal to estimate a class-wise correction for subsequent predictions.

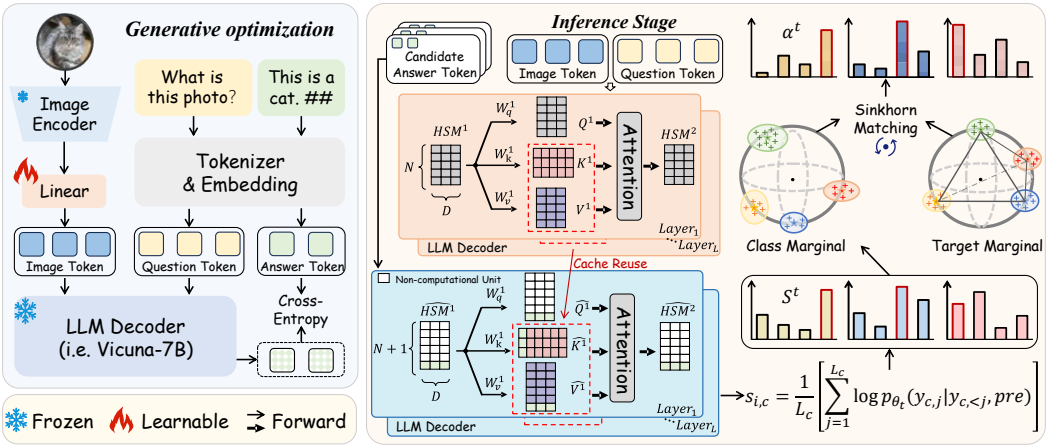


Figure 2: **Overview of UnL^2.** Generative optimization associates each accumulated class with a candidate answer sequence. UnL^2 converts free-form generation into generative evidence over the accumulated label space through Candidate Scoring, suppressing preference toward responses outside valid class candidates while exposing persistent generative preference among candidate classes. Online Calibration estimates a class-wise correction from generative evidence stored in a calibration queue through Sinkhorn matching, producing calibrated generative evidence for prediction. HSM represents the hidden state matrix of the layer's input in the model.

4.1 Candidate Scoring for Generative Evidence

Candidate Scoring evaluates the candidate answer sequence of every accumulated class under the same image-question context, transforming free-form generation into comparable generative evidence over $\mathcal{C}_t$ while suppressing preference toward responses outside the accumulated label space. Since candidate answer sequences share the image $\mathbf{x}_i^t$ and question $\mathbf{q}$ and differ only in their answer tokens, UnL^2 reuses the shared multimodal prefix across all accumulated classes rather than recomputing it for each candidate. Let $\mathbf{H}_{\text{pre}}$ denote the hidden states of the shared image-question prefix, with cached key-value states $\mathbf{K}_{\text{pre}} = \mathbf{H}_{\text{pre}}\mathbf{W}_K$ and $\mathbf{V}_{\text{pre}} = \mathbf{H}_{\text{pre}}\mathbf{W}_V$. When evaluating the ℓ-th token of candidate answer sequence $\mathbf{a}^c$, its hidden state $\mathbf{h}_\ell^c$ produces $\mathbf{q}_\ell^c = \mathbf{h}_\ell^c\mathbf{W}_Q$, $\mathbf{k}_\ell^c = \mathbf{h}_\ell^c\mathbf{W}_K$, and $\mathbf{v}_\ell^c = \mathbf{h}_\ell^c\mathbf{W}_V$. Attention is computed over the concatenated key-value states $[\mathbf{K}_{\text{pre}}, \mathbf{K}_{<\ell}^c]$ and $[\mathbf{V}_{\text{pre}}, \mathbf{V}_{<\ell}^c]$, where $\mathbf{K}_{<\ell}^c$ and $\mathbf{V}_{<\ell}^c$ contain the states generated by preceding answer tokens. Newly generated $\mathbf{k}_\ell^c$ and $\mathbf{v}_\ell^c$ are appended to the candidate-specific cache for subsequent decoding (Ge et al., 2024; Qin et al., 2025; Tu et al., 2025). Reusing $\mathbf{K}_{\text{pre}}$ and $\mathbf{V}_{\text{pre}}$ preserves the original autoregressive likelihood while amortizing multimodal computation across the K_t accumulated classes. For each candidate class $c \in \mathcal{C}_t$, UnL^2 defines the generative score as the length-normalized autoregressive log-likelihood of its candidate answer sequence,

$$s_{i,c}^t = \frac{1}{L_c}\sum_{\ell=1}^{L_c} \log p_{\boldsymbol{\theta}^t}\left(a_\ell^c \mid \mathbf{x}_i^t, \mathbf{q}, \mathbf{a}_{<\ell}^c\right), \tag{2}$$

where normalization by the answer length L_c prevents sequence length from confounding comparisons across candidate classes. Aggregating all generative scores yields the generative evidence $\mathbf{s}_i^t = [s_{i,c}^t]_{c\in\mathcal{C}_t} \in \mathbb{R}^{K_t}$, which represents the relative generative support over the entire accumulated label space. Confining evaluation to valid candidate classes suppresses preference outside the accumulated label space while exposing generative preference among the accumulated classes in $\mathbf{s}_i^t$.

4.2 Online Calibration with Sinkhorn Matching

Persistent generative preference contains a class-dependent component shared across samples, whereas visual evidence is sample-specific. Online Calibration estimates this shared component from a bounded FIFO calibration queue $\mathcal{M}^t$ containing generative evidence from previously processed test samples (Hou et al., 2026; Huang & Zhu, 2026). At evaluation after each incremental task t, $\mathcal{M}^t$ is reset to empty and samples are processed sequentially over the accumulated label space $\mathcal{C}_t$. Each generative evidence vector is enqueued only after its prediction, with FIFO replacement once the capacity M is reached, ensuring that calibration of $\mathbf{x}_i^t$ depends only on preceding samples in the current evaluation (Jo et al., 2025). When $\mathcal{M}^t$ is empty, we set $\boldsymbol{\alpha}^t = \mathbf{0}$ and prediction reduces to uncalibrated Candidate Scoring. For $m = |\mathcal{M}^t| > 0$, let $\mathbf{S}^t \in \mathbb{R}^{m \times K_t}$ denote the queued generative evidence matrix. Cross-sample aggregation suppresses sample-specific variation while preserving persistent class-dependent preference, enabling Online Calibration to estimate a class-wise correction by aligning the predictive class marginal with a target marginal $\boldsymbol{\pi}^t \in \mathbb{R}^{K_t}$. We use the uniform target $\pi_c^t = 1/K_t$ as a distribution-independent calibration reference.

Let $\mathbf{Q}^t \in \mathbb{R}_+^{m \times K_t}$ denote the predictive assignment induced by the queued generative evidence over accumulated classes, with class marginal $\frac{1}{m}(\mathbf{Q}^t)^\top \mathbf{1}_m$. Sinkhorn matching (Cuturi, 2013) obtains the optimal assignment by matching this marginal to $\boldsymbol{\pi}^t$:

$$\mathbf{Q}^{t*} = \arg\max_{\mathbf{Q} \geq 0} \left\langle \mathbf{Q}, \mathbf{S}^t \right\rangle + \tau H(\mathbf{Q}) \quad \text{s.t.} \quad \mathbf{Q}\mathbf{1}_{K_t} = \mathbf{1}_m, \qquad \frac{1}{m}\mathbf{Q}^\top \mathbf{1}_m = \boldsymbol{\pi}^t, \tag{3}$$

where $H(\mathbf{Q}) = -\sum_{r,c} Q_{r,c} \log Q_{r,c}$ is the assignment entropy and τ is the entropy regularization coefficient. The row constraint normalizes the predictive assignment for each queued sample, while the column constraint aligns the aggregate class marginal with $\boldsymbol{\pi}^t$. Sinkhorn scaling induces a class-wise correction $\boldsymbol{\alpha}^t = [\alpha_c^t]_{c \in \mathcal{C}_t}$ satisfying $Q_{r,c}^{t*} \propto \exp((S_{r,c}^t - \alpha_c^t)/\tau)$, so classes with stronger persistent generative preference receive larger corrections. The calibrated evidence for the current sample is $\mathbf{s}_i^t - \boldsymbol{\alpha}^t$, yielding the prediction $\hat{y}_i^t = \arg\max_{c \in \mathcal{C}_t}(s_{i,c}^t - \alpha_c^t)$.

4.3 Theoretical Analysis

To characterize the stability of class-wise corrections estimated from queued generative evidence when applied to the current sample, we show that Sinkhorn matching exactly absorbs the class-wise component of generative preference shared across samples, while sample-specific deviations induce only bounded perturbations in calibrated generative evidence. Motivated by the empirically observed cross-sample persistence of generative preference, we model the perturbed queue evidence as $\overline{\mathbf{S}}^t = \mathbf{S}^t + \mathbf{1}_m(\mathbf{d}^t)^\top + \mathbf{E}^t$ and the perturbed evidence of the current sample as $\overline{\mathbf{s}}_i^t = \mathbf{s}_i^t + \mathbf{d}^t + \mathbf{e}_i^t$, where $\mathbf{d}^t$ represents shared class-wise preference and $\mathbf{E}^t$ and $\mathbf{e}_i^t$ capture sample-specific residuals satisfying $\|\mathbf{E}^t\|_\infty \leq \epsilon$ and $\|\mathbf{e}_i^t\|_\infty \leq \epsilon$. Let $\mathbf{Q}^{t*}$ and $\overline{\mathbf{Q}}^{t*}$ denote the optimal predictive assignments induced by $\mathbf{S}^t$ and $\overline{\mathbf{S}}^t$ under the same target marginal $\boldsymbol{\pi}^t$ and entropy coefficient $\tau > 0$, with corresponding class-wise corrections $\boldsymbol{\alpha}^t$ and $\overline{\boldsymbol{\alpha}}^t$. Accounting for the class-independent ambiguity of the Sinkhorn dual variables, let $\gamma^t > 0$ denote the minimum dual curvature on the subspace orthogonal to $\mathbf{1}_{K_t}$ along the perturbation path, characterizing the conditioning of class-wise calibration.

Theorem 1. *For any task t, suppose $\pi_c^t > 0$ for all $c \in \mathcal{C}_t$ and the Sinkhorn dual has curvature at least γ^t/τ on the subspace orthogonal to $\mathbf{1}_{K_t}$. Under the bounded residuals defined above, the deviation of calibrated pairwise generative evidence satisfies*

$$\begin{aligned} &\max_{c,c' \in \mathcal{C}_t} \left| \left[(\overline{s}_{i,c}^t - \overline{\alpha}_c^t) - (\overline{s}_{i,c'}^t - \overline{\alpha}_{c'}^t) \right] - \left[(s_{i,c}^t - \alpha_c^t) - (s_{i,c'}^t - \alpha_{c'}^t) \right] \right| \\ &\qquad \leq 2\|\mathbf{e}_i^t\|_\infty + \frac{\sqrt{K_t}}{\gamma^t}\|\mathbf{E}^t\|_\infty \leq \left(2 + \frac{\sqrt{K_t}}{\gamma^t}\right)\epsilon. \end{aligned} \tag{4}$$

Consequently, the shared class-wise preference $\mathbf{d}^t$ is eliminated from the calibration error, while deviations from the shared structure affect calibrated generative evidence only through sample-specific residuals. When $\epsilon = 0$, exact invariance is recovered with $\overline{\mathbf{Q}}^{t*} = \mathbf{Q}^{t*}$ and $\overline{\boldsymbol{\alpha}}^t = \boldsymbol{\alpha}^t + \mathbf{d}^t + \kappa^t \mathbf{1}_{K_t}$ for some class-independent offset κ^t.

As established in Theorem 1, shared class-wise preference is absorbed without altering the optimal predictive assignment, whereas calibration error is governed by sample-specific deviations. Under

the target marginal, $\mathbf{1}_m(\mathbf{d}^t)^\top$ contributes the constant $m(\boldsymbol{\pi}^t)^\top\mathbf{d}^t$ to every feasible Sinkhorn objective, leaving the optimum unchanged and allowing $\mathbf{d}^t$ to be absorbed into the class-wise correction. Sample-specific residuals perturb the assignment and correction, with entropic regularization yielding $\frac{1}{m}\|\overline{\mathbf{Q}}^{t*} - \mathbf{Q}^{t*}\|_1 \le \|\mathbf{E}^t\|_\infty/\tau$ and dual curvature yielding $\min_{\kappa^t} \|\overline{\boldsymbol{\alpha}}^t - \boldsymbol{\alpha}^t - \mathbf{d}^t - \kappa^t\mathbf{1}_{K_t}\|_2 \le \sqrt{K_t}\|\mathbf{E}^t\|_\infty/(2\gamma^t)$. Combining the correction perturbation induced by $\mathbf{E}^t$ with the current sample residual $\mathbf{e}_i^t$ gives Eq. (4), establishing that pairwise calibrated evidence varies linearly with non-shared deviations and inversely with the conditioning of Sinkhorn calibration. Consequently, if Δ_i^t denotes the margin between the largest and second-largest calibrated generative evidence for $\mathbf{x}_i^t$, the prediction remains unchanged whenever $\Delta_i^t > (2 + \sqrt{K_t}/\gamma^t)\epsilon$. Detailed proofs are in Appendix A.

Table 1: Performance results on Tiny-ImageNet. "Avg" represents the averaged performance after training each task, and "Last" represents the performance on all test samples after training the last task. * indicates our re-implementation based on original github repository.

Type	Method	Exemplar	Tiny-ImageNet					
			Inc. 5 tasks		Inc. 10 tasks		Inc. 20 tasks	
			Avg	**Last**	**Avg**	**Last**	**Avg**	**Last**
Conventional	iCaRL	✓	45.95	34.60	43.22	33.22	37.85	27.54
	EEIL	✓	47.17	35.12	45.03	34.64	40.41	29.72
	UCIR	✓	50.30	39.42	48.58	37.29	42.84	30.85
	PASS	✗	49.54	41.64	47.19	39.27	42.01	32.93
	DyTox	✓	55.58	47.23	52.26	42.79	46.18	36.21
Discriminative PT models	Continual-CLIP	✗	70.49	66.43	70.55	66.43	70.51	66.43
	DualPrompt	✗	85.15	81.01	81.38	73.73	73.45	60.16
	CODA-Prompt	✗	85.91	81.36	82.80	75.28	77.43	66.32
	MoE-CLIP	✗	81.12	76.81	80.23	76.35	79.96	75.77
	RAPF	✗	78.64	74.67	77.42	73.57	76.29	72.65
Generative PT models	Zero-shot	✗	58.16	53.72	58.10	53.72	58.13	53.72
	GMM	✗	83.42	76.98	82.49	76.51	81.70	76.03
	GMM*	✗	81.54	77.30	80.40	76.49	80.00	76.32
	KG-GMM	✗	86.17	81.86	84.37	78.16	83.18	78.32
	UnL^2 (Ours)	✗	**87.51** ±0.19	**82.43** ±0.68	**85.40** ±0.12	**82.58** ±0.61	**84.06** ±0.08	**82.38** ±0.66

5 EXPERIMENTS

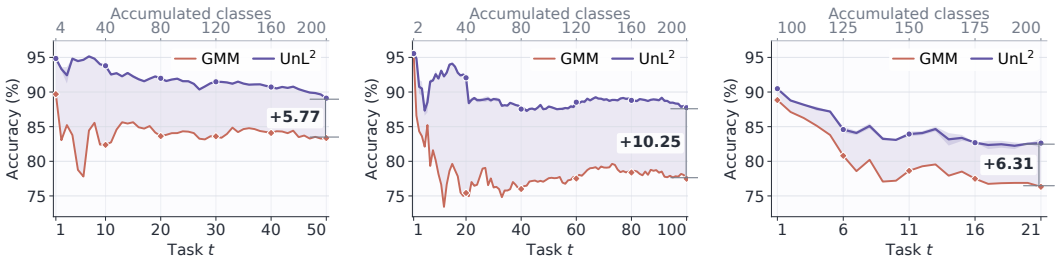
5.1 EXPERIMENTAL SETUPS

Datasets and Baselines. We evaluate on Tiny-ImageNet (Le & Yang, 2015) and ImageNet-R (Hendrycks et al., 2021), both containing 200 classes. Following prior generative CIL protocols (Cao et al., 2024; 2025), Tiny-ImageNet adopts the B100 setting, where the first session contains 100 classes and the remaining classes are divided into 5, 10, or 20 incremental sessions. For ImageNet-R, we adopt the B0 setting and evenly partition all classes into 10, 20, 25, 50, or 100 tasks, allowing evaluation over increasingly long incremental sequences. We compare with representative methods from three categories: conventional CIL methods, including iCaRL (Rebuffi et al., 2017), EEIL (Castro et al., 2018), UCIR (Hou et al., 2019), PASS (Zhu et al., 2021), and DyTox (Douillard et al., 2022); discriminative PTM-based methods, including Continual-CLIP (Thengane et al., 2022), DualPrompt (Wang et al., 2022), CODA-Prompt (Smith et al., 2023), MoE-CLIP (Yu et al., 2024a), RAPF (Huang et al., 2024), InfLoRA (Liang & Li, 2024), EASE (Zhou et al., 2024a), MOS (Sun et al., 2025), MACIL (Wu et al., 2025), CoSO (Cheng et al., 2025), and EWC-LoRA (Zheng et al., 2026); and generative methods, including GMM (Cao et al., 2024), and KG-GMM (Cao et al., 2025). As KG-GMM has no available implementation, we report its published results where available.

Implementation Details. Following (Cao et al., 2024), we adopt MiniGPT-4 (Zhu et al., 2024) with EVA-CLIP ViT-G/14 (Fang et al., 2023), BLIP-2 Q-Former (Li et al., 2023b), and Vicuna-7B (Chiang et al., 2023). We optimize only the vision-to-language projection initialized from the pretrained MiniGPT-4 checkpoint, while keeping all other components frozen. On ImageNet-R (B0), the projection is optimized for two epochs per task with a learning rate of 3×10^{-7} and cosine decay. On Tiny-ImageNet (B100), the learning rate is 3×10^{-6} for the initial 100-class session and 3×10^{-7} thereafter, with the same cosine decay. More details are provided in the Appendix B.

Table 2: Performance results on ImageNet-R. "Avg" represents the averaged performance after training each task, and "Last" represents the performance on all test samples after training the last task.

Method	ImageNet-R									
	10 tasks		20 tasks		25 tasks		50 tasks		100 tasks	
	Avg	**Last**	**Avg**	**Last**	**Avg**	**Last**	**Avg**	**Last**	**Avg**	**Last**
CODA-Prompt	78.47	73.31	75.34	69.96	70.47	63.86	55.59	48.89	38.71	30.92
InfLoRA	81.90	75.88	79.76	73.01	76.84	69.09	71.20	61.91	51.56	41.73
EASE	81.38	75.91	80.28	73.78	79.65	72.69	75.49	68.53	58.33	50.37
MOS	81.94	77.65	80.39	75.04	79.43	73.66	74.81	66.92	56.19	47.18
MACIL	85.76	81.82	84.25	79.46	81.90	76.18	77.47	70.10	63.04	55.02
CoSO	85.56	81.10	82.77	76.68	81.74	74.98	73.03	64.75	54.34	45.83
EWC-LoRA	78.95	72.86	69.71	66.85	68.00	63.37	61.14	46.59	39.28	27.81
Zero-shot	70.58	65.67	71.14	65.67	71.17	65.67	71.51	65.67	71.65	65.67
GMM	84.62	84.37	84.97	84.13	84.53	84.00	83.98	83.35	78.19	77.48
UnL^2 (Ours)	**91.40** ±0.05	**89.43** ±0.08	**91.26** ±0.02	**89.07** ±0.08	**91.13** ±0.01	**88.87** ±0.10	**90.42** ±0.02	**88.68** ±0.03	**87.00** ±0.01	**87.05** ±0.25



(a) ImageNet-R, 50 Tasks (b) ImageNet-R, 100 Tasks (c) Tiny-ImageNet, 21 Tasks

Figure 3: The performance of each learning session under different settings of ImageNet-R and Tiny-ImageNet. UnL^2 consistently outperforms Cao et al. (2024) throughout the process, with the performance advantage becoming more pronounced as the incremental sequence length increases.

5.2 Main Results

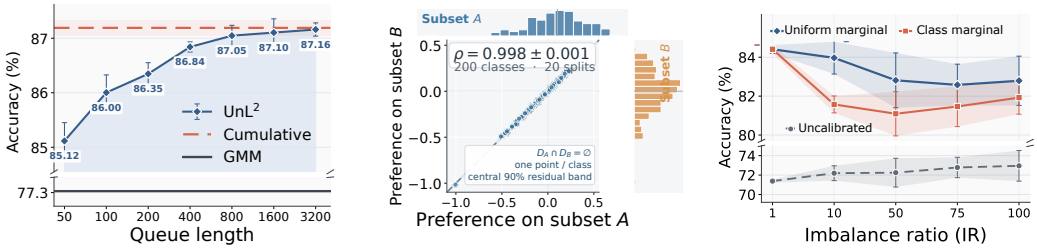
As shown in Tables 1 and 2, UnL^2 consistently outperforms both discriminative CIL methods and existing generative approaches across all incremental settings on Tiny-ImageNet and ImageNet-R, with increasingly pronounced gains as the number of incremental tasks grows. Larger improvements under longer incremental sequences align with the increasing inference interference induced by label-space expansion, as more accumulated classes amplify the effect of persistent generative preference on prediction. By deriving comparable generative evidence over the accumulated label space and calibrating persistent generative preference among candidate classes, UnL^2 remains robust to the increasing competition introduced by accumulated classes. Task-wise results in Figure 3 further reveal UnL^2 continuously mitigates inference interference throughout the incremental process rather than recovering performance only at the final stage.

5.3 Understanding Unlocking Label Likelihoods

Impact of Each Component. Table 3 disentangles the methodological contributions of UnL^2 from the effect of historical information access. Equipping GMM with the same history-based calibration protocol improves its performance but remains consistently below UnL^2, showing that the gains cannot be explained by access to previously processed samples alone. Candidate Scoring replaces free-form generation with comparable class-wise generative evidence over valid candidate classes, thereby eliminating generative preference toward responses outside the accumulated label space from prediction. Comparing the full UnL^2 with its variant without Online Calibration further shows that correcting persistent generative preference among candidate classes yields increasingly pronounced gains as the accumulated label space expands. Together, these results show that UnL^2 addresses complementary sources of generative preference through Candidate Scoring and Online Calibration, rather than benefiting merely from historical information access. Additional task-wise accuracy curves across multiple settings are provided in Figure 9.

Table 3: Ablation study on ImageNet-R with different tasks settings. "Avg" denotes the average performance across all tasks, while "Last" denotes the performance over the all label space.

Method	ImageNet-R									
	10 tasks		20 tasks		25 tasks		50 tasks		100 tasks	
	Avg	**Last**	**Avg**	**Last**	**Avg**	**Last**	**Avg**	**Last**	**Avg**	**Last**
GMM	84.62	84.37	84.97	84.13	84.53	84.00	83.98	83.35	78.19	77.48
GMM w/ OC	85.41	84.90	85.35	84.93	85.35	84.91	84.67	84.35	79.74	79.24
UnL^2 w/o OC	86.85	86.98	86.61	86.00	86.23	85.27	84.29	84.87	68.94	77.33
UnL^2	**91.40**	**89.43**	**91.26**	**89.07**	**91.13**	**88.87**	**90.42**	**88.68**	**87.00**	**87.05**



(a) Calibration Queue (b) Generative Preference (c) Target Marginal

Figure 4: Analysis of Online Calibration. (a) Increasing the calibration queue capacity improves the estimation of class-wise corrections, with performance progressively approaching cumulative calibration. (b) Generative preference estimated from disjoint sample subsets remains highly consistent, supporting its persistence across samples. (c) Under increasing class imbalance, the uniform target marginal consistently outperforms the oracle class marginal.

Stability under Preference Perturbations. Figure 4(b) supports the cross-sample persistence underlying Theorem 1, as generative preference estimated from disjoint sample subsets remains highly consistent across accumulated classes. Following the perturbation model in Theorem 1, we independently vary the strength λ of shared class-wise preference and the bound ϵ on sample-specific residuals, re-estimating the class-wise correction under each perturbation setting. Figure 5 shows that UnL^2 preserves complete prediction agreement as λ increases when $\epsilon = 0$, whereas uncalibrated predictions deviate substantially from the unperturbed case. With the shared preference fixed, increasing ϵ leads to a gradual reduction in prediction agreement, consistent with the bounded sensitivity of calibrated generative evidence to sample-specific residuals established in Theorem 1. Together, these results distinguish exact invariance to shared class-wise preference from bounded degradation induced by sample-specific residuals.

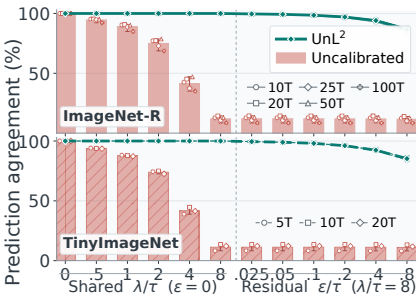


Figure 5: Prediction agreement under controlled preference perturbations.

Impact of Calibration Queue. Figure 4(a) examines the effect of calibration queue capacity M on class-wise correction estimation. Increasing M broadens the cross-sample coverage of queued generative evidence, reducing variation in the estimated predictive class marginal and yielding progressively more stable corrections. Performance approaches cumulative calibration as the capacity increases, indicating that persistent generative preference can be reliably characterized using a bounded queue without retaining generative evidence from all previously processed samples.

Effect of Target Marginal. Figure 4(c) examines the effect of the target marginal under increasing class imbalance by comparing the uniform target with the oracle class marginal. The uniform target consistently yields higher accuracy, supporting its role as a class-symmetric reference for estimating persistent generative preference rather than modeling the underlying class distribution. Since the correction targets model-induced class-wise preference, incorporating the oracle marginal introduces distribution-dependent asymmetry into Sinkhorn matching, whereas the uniform target preserves a common class-symmetric calibration reference across distributions.

5.4 Further Analysis

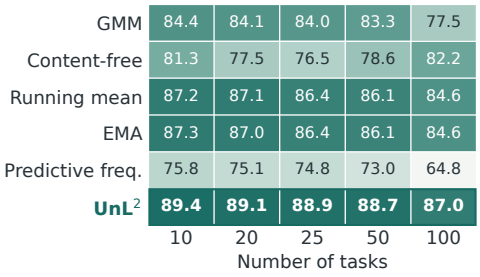


Figure 6: Comparison of different calibration strategies on ImageNet-R.

Calibration Strategy. Figure 6 compares Online Calibration with different calibration methods across different task settings. Content-free calibration estimates class-wise preference from a null visual reference (Zhao et al., 2021; Holtzman et al., 2021), but its inconsistent performance across settings indicates that a single reference cannot reliably characterize generative preference. History-based calibration methods use the same generative evidence from previously processed samples as UnL^2 and estimate class-wise corrections using the running mean, exponential moving average (EMA), or predictive frequency. Their consistently inferior performance shows that historical information access alone does not account for the gains and that simple summary statistics are insufficient to characterize generative preference across samples. In contrast, UnL^2 performs Sinkhorn matching over queued generative evidence to align the predictive class marginal with the target marginal, yielding more effective class-wise corrections.

Backbone Generalization & Efficiency. Table 4 shows that UnL^2 consistently outperforms GMM on both Qwen2.5-VL-7B (Bai et al., 2025) and LLaVA-OneVision-7B (Li et al., 2025), demonstrating its generalization across MLLM backbones. The efficient candidate evaluation in Candidate Scoring achieves $27.42\times$ and $33.84\times$ speedups over sequential evaluation, respectively.

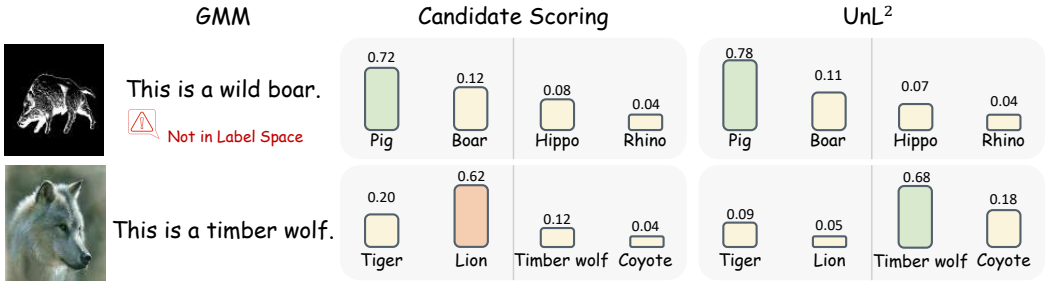


Figure 7: Qualitative analysis of UnL^2.

Qualitative Analysis. Figure 7 illustrates two representative failure modes addressed by UnL^2. For *pig*, GMM generates the out-of-label-space response "wild boar", whereas Candidate Scoring restricts evaluation to accumulated candidate classes and recovers the correct prediction. For *timber wolf*, Candidate Scoring assigns the highest generative score to *lion*, revealing persistent generative preference within the accumulated label space, while Online Calibration reduces this preference through the estimated class-wise correction and restores the correct prediction.

Table 4: **Backbone generalization and efficiency** on ImageNet-R with 10 tasks. "Sequential" evaluates candidate likelihoods without the efficient design in Candidate Scoring.

	Qwen2.5-VL-7B	LLaVA-OneVision-7B
	Last Accuracy (%)	
GMM	82.78	88.17
UnL^2	**86.77**	**91.78**
	Latency (s/sample)	
GMM	0.820	0.520
Sequential.	23.771	49.920
UnL^2	0.867	1.475
Speedup	**27.42**×	**33.84**×

6 Conclusion

We identify an overlooked inference interference induced by label-space expansion in generative CIL, where persistent generative preference among candidate classes increasingly biases predictions toward newly introduced classes and degrades old-class recognition. UnL^2 mitigates this interference by converting free-form generation into comparable class-wise generative evidence over the accumulated label space through Candidate Scoring and calibrating persistent generative preference with a class-wise correction estimated by Online Calibration via Sinkhorn matching. Theoretical analysis shows that Sinkhorn calibration exactly absorbs the shared additive class-wise preference component while bounding the effect of sample-specific deviations on calibrated pairwise evidence. Experiments across multiple CIL benchmarks consistently demonstrate improvements over the evaluated baselines, with increasingly pronounced gains under longer incremental sequences.

AI Assistance Disclosure

The authors used generative AI tools solely for language editing and proofreading to improve the clarity and readability of the manuscript. All scientific work, including the conception of ideas, development of methods, conduct of experiments, analysis and interpretation of results, and formulation of conclusions, was carried out by the authors. Generative AI tools were not used to generate scientific ideas, methods, data, analyses, or conclusions.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, 2025.

Xusheng Cao, Haori Lu, Linlan Huang, Xialei Liu, and Ming-Ming Cheng. Generative multi-modal models are good class-incremental learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28706–28717, 2024.

Xusheng Cao, Haori Lu, Linlan Huang, Fei Yang, Xialei Liu, and Ming-Ming Cheng. Knowledge graph enhanced generative multi-modal models for class-incremental learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *European Conference on Computer Vision (ECCV)*, pp. 233–248, 2018.

Hanzhu Chen, Xu Shen, Jie Wang, Zehao Wang, Qitan Lv, Junjie He, Rong Wu, Feng Wu, and Jieping Ye. Knowledge graph finetuning enhances knowledge manipulation in large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Quan Cheng, Yuanyu Wan, Lingyu Wu, Chenping Hou, and Lijun Zhang. Continuous subspace optimization for continual learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality, 2023.

Seungho Choi, Sihyun Park, Minsang Kim, Chansol Park, and Bongsu Kim. Hanjabridge: Resolving semantic ambiguity in korean llms via hanja-augmented pre-training. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 30413–30421, 2026.

Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. In *Conference on Neural Information Processing Systems (NeurIPS)*, pp. 58921–58937, 2023.

Alessandro Conti, Massimiliano Mancini, Enrico Fini, Yiming Wang, Paolo Rota, and Elisa Ricci. On large multimodal models as open-world image classifiers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16388–16398, 2025.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Conference on Neural Information Processing Systems (NeurIPS)*, pp. 2292–2300, 2013.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Conference on Neural Information Processing Systems (NeurIPS)*, pp. 49250–49267, 2023.

Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9285–9295, 2022.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the limits of masked visual representation learning at scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19358–19369, 2023.

Mingtao Feng, Fenghao Tian, Jianqiao Luo, Zijie Wu, Weisheng Dong, Yaonan Wang, and Ajmal Saeed Mian. Semantic ambiguity modeling and propagation for fine-grained visual cross view geo-localization. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 2978–2986, 2025.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive KV cache compression for llms. In *International Conference on Learning Representations (ICLR)*, 2024.

Qiushan Guo, Chuofan Ma, Yi Jiang, Zehuan Yuan, Yizhou Yu, and Ping Luo. EGC: Image generation and classification via a diffusion energy-based model. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22952–22962, 2023.

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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349, 2021.

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

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7038–7051, 2021.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 831–839, 2019.

Yue Hou, Ruomei Liu, Yingke Su, Junran Wu, and Ke Xu. Graph out-of-distribution detection via test-time calibration with dual dynamic dictionaries. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

Linlan Huang, Xusheng Cao, Haori Lu, and Xialei Liu. Class-incremental learning with CLIP: Adaptive representation adjustment and parameter fusion. In *European Conference on Computer Vision (ECCV)*, pp. 214–231, 2024.

Linlan Huang, Xusheng Cao, Haori Lu, Yifan Meng, Fei Yang, and Xialei Liu. Mind the gap: Preserving and compensating for the modality gap in clip-based continual learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3777–3786, 2025.

Qunjie Huang and Weina Zhu. Sattc: Structure-aware label-free test-time calibration for cross-subject eeg-to-image retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Runhui Huang, Jianhua Han, Guansong Lu, Xiaodan Liang, Yihan Zeng, Wei Zhang, and Hang Xu. DiffDis: Empowering generative diffusion model with cross-modal discrimination capability. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15713–15723, 2023.

Yong-Yeon Jo, Byeong Tak Lee, Jeong-Ho Hong, Hak Seung Lee, Joon myoung Kwon, and Beom Joon Kim. Test-time calibration: A framework for personalized test-time adaptation in real-world biosignals. In *Conference on Health, Inference, and Learning (CHIL)*, 2025.

Logan Lawrence, Oindrila Saha, Megan Wei, Chen Sun, Subhransu Maji, and Grant Van Horn. You may speak freely: Improving the fine-grained visual recognition capabilities of multimodal large language models with answer extraction. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1428–1437, 2026.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 2015.

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2206–2217, 2023a.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, pp. 19730–19742, 2023b.

Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23638–23647, 2024.

Wenzhuo Liu, Fei Zhu, Longhui Wei, and Qi Tian. C-CLIP: multimodal continual learning for vision-language model. In *International Conference on Learning Representations (ICLR)*, 2025.

Xialei Liu, Marc Masana, Luis Herranz, Joost Van de Weijer, Antonio M Lopez, and Andrew D Bagdanov. Rotate your networks: Better weight consolidation and less catastrophic forgetting. In *International Conference on Pattern Recognition (ICPR)*, pp. 2262–2268, 2018.

Xuan Liu and Xiaobin Chang. Lora subtraction for drift-resistant space in exemplar-free continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15308–15318, 2025.

Tengfei Ma, Xiang Song, Wen Tao, Mufei Li, Jiani Zhang, Xiaoqin Pan, Yijun Wang, Bosheng Song, and Xiangxiang Zeng. Towards synergistic path-based explanations for knowledge graph completion: Exploration and evaluation. In *International Conference on Learning Representations (ICLR)*, 2025.

Youngmin Oh, Donghyeon Baek, and Bumsub Ham. Alife: Adaptive logit regularizer and feature replay for incremental semantic segmentation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Ziran Qin, Yuchen Cao, Mingbao Lin, Wen Hu, Shixuan Fan, Ke Cheng, Weiyao Lin, and Jianguo Li. CAKE: cascading and adaptive KV cache eviction with layer preferences. In *International Conference on Learning Representations (ICLR)*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2001–2010, 2017.

Mohammad Rostami, Soheil Kolouri, and Praveen K Pilly. Complementary learning for overcoming catastrophic forgetting using experience replay. *CoRR*, 2019.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Conference on Neural Information Processing Systems (NeurIPS)*, pp. 2990–2999, 2017.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11909–11919, 2023.

Hai-Long Sun, Da-Wei Zhou, Hanbin Zhao, Le Gan, De-Chuan Zhan, and Han-Jia Ye. MOS: Model surgery for pre-trained model-based class-incremental learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 20699–20707, 2025.

Vishal Thengane, Salman Khan, Munawar Hayat, and Fahad Khan. Clip model is an efficient continual learner. *CoRR*, 2022.

Dezhan Tu, Danylo Vashchilenko, Yuzhe Lu, and Panpan Xu. Vl-cache: Sparsity and modality-aware KV cache compression for vision-language model inference acceleration. In *International Conference on Learning Representations (ICLR)*, 2025.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *CoRR*, 2023.

Yan Wang, Da-Wei Zhou, and Han-Jia Ye. Integrating task-specific and universal adapters for pre-trained model-based class-incremental learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 806–816, 2025.

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 *European Conference on Computer Vision (ECCV)*, pp. 631–648, 2022.

Fangwen Wu, Lechao Cheng, Shengeng Tang, Xiaofeng Zhu, Chaowei Fang, Dingwen Zhang, and Meng Wang. Navigating semantic drift in task-agnostic class-incremental learning. In *International Conference on Machine Learning (ICML)*, pp. 67172–67183, 2025.

Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23219–23230, 2024a.

Qihang Yu, Xiaohui Shen, and Liang-Chieh Chen. Towards open-ended visual recognition with large language models. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *European Conference on Computer Vision (ECCV)*, volume 15072 of *Lecture Notes in Computer Science*, pp. 359–376, 2024b.

Huan Zhang, Shuyu Dong, Yujin Zheng, Dingwen Wang, Shenghua Fan, and Fan Lyu. Subspace alignment for clip-based continual learning via canonical correlation analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32257–32266, 2026.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning (ICML)*, pp. 12697–12706, 2021.

Yaoyue Zheng, Yin Zhang, Joost van de Weijer, Gido M. van de Ven, Shaoyi Du, Xuetao Zhang, and Zhiqiang Tian. Revisiting weight regularization for low-rank continual learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23554–23564, 2024a.

Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 9851–9873, 2024b.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5871–5880, 2021.

Hao Zhu, Yifei Zhang, Junhao Dong, and Piotr Koniusz. Bilora: Almost-orthogonal parameter spaces for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25613–25622, 2025.

A Detailed Proof

Proof of Theorem 1. For conciseness, we omit the task superscript t when no ambiguity arises and restore it in the final results. Throughout the proof, $\|\mathbf{E}\|_\infty = \max_{r,c} |E_{r,c}|$ denotes the entry-wise infinity norm, while $\|\mathbf{A}\|_1 = \sum_{r,c} |A_{r,c}|$ denotes the entry-wise ℓ_1 norm. Let

$$\mathcal{U} = \left\{ \mathbf{Q} \in \mathbb{R}_+^{m \times K_t} \;\middle|\; \mathbf{Q}\mathbf{1}_{K_t} = \mathbf{1}_m, \frac{1}{m}\mathbf{Q}^\top \mathbf{1}_m = \boldsymbol{\pi}^t \right\}$$

denote the feasible set of predictive assignments. For every $\mathbf{Q} \in \mathcal{U}$, the shared class-wise preference component satisfies

$$\left\langle \mathbf{Q}, \mathbf{1}_m (\mathbf{d}^t)^\top \right\rangle = (\mathbf{d}^t)^\top \mathbf{Q}^\top \mathbf{1}_m = m(\boldsymbol{\pi}^t)^\top \mathbf{d}^t.$$

Hence, $\mathbf{1}_m(\mathbf{d}^t)^\top$ contributes the same constant to the Sinkhorn objective for every feasible predictive assignment. Adding this shared component therefore leaves the relative objective values and the optimal predictive assignment unchanged. All deviations from exact invariance are consequently induced by the sample-specific residual $\mathbf{E}^t$.

We first quantify the perturbation of the optimal predictive assignment. Let $\mathbf{Q}^{t*}$ denote the unique optimum associated with $\mathbf{S}^t$ and let $\mathbf{Q}_E^{t*}$ denote the unique optimum associated with $\mathbf{S}^t + \mathbf{E}^t$. Strict concavity of the entropy-regularized objective over the convex feasible set guarantees uniqueness of both assignments. Since the shared component changes every feasible objective by the same constant, the optimum associated with $\overline{\mathbf{S}}^t = \mathbf{S}^t + \mathbf{1}_m(\mathbf{d}^t)^\top + \mathbf{E}^t$ satisfies $\overline{\mathbf{Q}}^{t*} = \mathbf{Q}_E^{t*}$.

Optimality of $\mathbf{Q}^{t*}$ and $\mathbf{Q}_E^{t*}$ yields corresponding row and column dual variables. Subtracting their first-order stationarity conditions and taking the inner product with $\mathbf{Q}_E^{t*} - \mathbf{Q}^{t*}$ eliminates all row- and column-dual terms because both assignments have identical marginals. This gives

$$\begin{aligned} \left\langle \mathbf{E}^t, \mathbf{Q}_E^{t*} - \mathbf{Q}^{t*} \right\rangle &= \tau \left\langle \log \mathbf{Q}_E^{t*} - \log \mathbf{Q}^{t*}, \mathbf{Q}_E^{t*} - \mathbf{Q}^{t*} \right\rangle \\ &= \tau \left[D_{\mathrm{KL}} \left(\mathbf{Q}_E^{t*} \,\middle\|\, \mathbf{Q}^{t*} \right) + D_{\mathrm{KL}} \left(\mathbf{Q}^{t*} \,\middle\|\, \mathbf{Q}_E^{t*} \right) \right], \end{aligned} \tag{5}$$

where $D_{\mathrm{KL}}(\mathbf{A}\|\mathbf{B}) = \sum_{r,c} A_{r,c} \log(A_{r,c}/B_{r,c})$. Each row of both assignments is a probability vector. Applying Pinsker's inequality to every row and then applying Cauchy–Schwarz across the m rows gives

$$\begin{aligned} &D_{\mathrm{KL}} \left(\mathbf{Q}_E^{t*} \,\middle\|\, \mathbf{Q}^{t*} \right) + D_{\mathrm{KL}} \left(\mathbf{Q}^{t*} \,\middle\|\, \mathbf{Q}_E^{t*} \right) \\ &\qquad \geq \sum_{r=1}^{m} \left\| \mathbf{Q}_{E,r}^{t*} - \mathbf{Q}_r^{t*} \right\|_1^2 \geq \frac{1}{m} \left\| \mathbf{Q}_E^{t*} - \mathbf{Q}^{t*} \right\|_1^2. \end{aligned} \tag{6}$$

Hölder's inequality also gives

$$\left\langle \mathbf{E}^t, \mathbf{Q}_E^{t*} - \mathbf{Q}^{t*} \right\rangle \leq \|\mathbf{E}^t\|_\infty \left\| \mathbf{Q}_E^{t*} - \mathbf{Q}^{t*} \right\|_1.$$

Combining Eq. (5) and Eq. (6) and cancelling the nonzero factor $\|\mathbf{Q}_E^{t*} - \mathbf{Q}^{t*}\|_1$ whenever the two assignments differ yields

$$\frac{1}{m} \left\| \overline{\mathbf{Q}}^{t*} - \mathbf{Q}^{t*} \right\|_1 \leq \frac{\|\mathbf{E}^t\|_\infty}{\tau}. \tag{7}$$

Equality of the assignments makes Eq. (7) immediate, so the bound holds in all cases.

We next quantify how $\mathbf{E}^t$ perturbs the class-wise correction. Introduce the interpolation $\mathbf{S}^t(u) = \mathbf{S}^t + u\mathbf{E}^t$ for $u \in [0,1]$ and let $\boldsymbol{\alpha}^t(u)$ denote the corresponding Sinkhorn correction. Because adding a constant to every component of $\boldsymbol{\alpha}^t(u)$ leaves the predictive assignment unchanged, we fix the gauge $\mathbf{1}_{K_t}^\top \boldsymbol{\alpha}^t(u) = 0$. For the r-th queued evidence vector, the corresponding predictive assignment is

$$q_{r,c}^t(u) = \frac{\exp\left((S_{r,c}^t + uE_{r,c}^t - \alpha_c^t(u))/\tau \right)}{\sum_{k \in \mathcal{C}_t} \exp\left((S_{r,k}^t + uE_{r,k}^t - \alpha_k^t(u))/\tau \right)}.$$

A normalized Sinkhorn dual associated with this interpolation is

$$\Phi_u(\boldsymbol{\alpha}) = \frac{\tau}{m} \sum_{r=1}^{m} \log \sum_{c \in \mathcal{C}_t} \exp\left(\frac{S_{r,c}^t + uE_{r,c}^t - \alpha_c}{\tau} \right) + (\boldsymbol{\pi}^t)^\top \boldsymbol{\alpha}.$$

Its first-order condition is

$$\nabla\Phi_u(\boldsymbol{\alpha}^t(u)) = \boldsymbol{\pi}^t - \frac{1}{m}\sum_{r=1}^{m}\mathbf{q}_r^t(u) = \mathbf{0},$$

which is precisely the target-marginal constraint. Define

$$\mathbf{C}_r^t(u) = \mathrm{Diag}\left(\mathbf{q}_r^t(u)\right) - \mathbf{q}_r^t(u)\mathbf{q}_r^t(u)^\top, \qquad \mathbf{C}^t(u) = \frac{1}{m}\sum_{r=1}^{m}\mathbf{C}_r^t(u).$$

Direct differentiation gives

$$\nabla^2\Phi_u\left(\boldsymbol{\alpha}^t(u)\right) = \frac{1}{\tau}\mathbf{C}^t(u).$$

Each $\mathbf{C}_r^t(u)$ is positive semidefinite and has $\mathbf{1}_{K_t}$ as its null direction. Strict positivity of the predictive probabilities makes $\mathbf{C}^t(u)$ positive definite on $\mathbf{1}_{K_t}^\perp$. Consistently with Theorem 1, define the minimum curvature along the interpolation by

$$\gamma^t = \inf_{u\in[0,1]} \lambda_{\min}\left(\mathbf{C}^t(u)\big|_{\mathbf{1}_{K_t}^\perp}\right) > 0. \tag{8}$$

Corresponding dual curvature is therefore bounded below by γ^t/τ on $\mathbf{1}_{K_t}^\perp$.

Differentiating the marginal constraint $\frac{1}{m}\sum_{r=1}^{m}\mathbf{q}_r^t(u) = \boldsymbol{\pi}^t$ with respect to u gives

$$\frac{1}{m\tau}\sum_{r=1}^{m}\mathbf{C}_r^t(u)\left(\mathbf{E}_r^t - \frac{\mathrm{d}\boldsymbol{\alpha}^t(u)}{\mathrm{d}u}\right) = \mathbf{0},$$

where $\mathbf{E}_r^t$ denotes the r-th row of $\mathbf{E}^t$. Rearranging gives

$$\mathbf{C}^t(u)\frac{\mathrm{d}\boldsymbol{\alpha}^t(u)}{\mathrm{d}u} = \frac{1}{m}\sum_{r=1}^{m}\mathbf{C}_r^t(u)\mathbf{E}_r^t. \tag{9}$$

Gauge normalization ensures $\frac{\mathrm{d}\boldsymbol{\alpha}^t(u)}{\mathrm{d}u} \in \mathbf{1}_{K_t}^\perp$, while the right-hand side of Eq. (9) also belongs to $\mathbf{1}_{K_t}^\perp$ because $\mathbf{1}_{K_t}^\top\mathbf{C}_r^t(u) = \mathbf{0}^\top$.

For any probability vector $\mathbf{q}$ and any vector $\mathbf{v}$ satisfying $\|\mathbf{v}\|_\infty \le \|\mathbf{E}^t\|_\infty$, the c-th component of $(\mathrm{Diag}(\mathbf{q}) - \mathbf{q}\mathbf{q}^\top)\mathbf{v}$ satisfies

$$\begin{aligned}\left|q_c\left(v_c - \mathbf{q}^\top\mathbf{v}\right)\right| &= \left|q_c\sum_{k\neq c} q_k(v_c - v_k)\right| \\ &\le 2q_c(1-q_c)\|\mathbf{E}^t\|_\infty \\ &\le \frac{1}{2}\|\mathbf{E}^t\|_\infty,\end{aligned}$$

where the last inequality follows from $q_c(1-q_c) \le 1/4$. Applying this bound to every coordinate of the right-hand side of Eq. (9) yields

$$\left\|\frac{1}{m}\sum_{r=1}^{m}\mathbf{C}_r^t(u)\mathbf{E}_r^t\right\|_2 \le \frac{\sqrt{K_t}}{2}\|\mathbf{E}^t\|_\infty.$$

Combining this result with Eq. (8) and Eq. (9) gives

$$\left\|\frac{\mathrm{d}\boldsymbol{\alpha}^t(u)}{\mathrm{d}u}\right\|_2 \le \frac{\sqrt{K_t}}{2\gamma^t}\|\mathbf{E}^t\|_\infty.$$

Integrating along $u \in [0,1]$ therefore gives

$$\left\|\boldsymbol{\alpha}^t(1) - \boldsymbol{\alpha}^t(0)\right\|_2 \le \frac{\sqrt{K_t}}{2\gamma^t}\|\mathbf{E}^t\|_\infty. \tag{10}$$

We now restore the shared class-wise preference $\mathbf{d}^t$. Because

$$\frac{\exp\left((S^t_{r,c} + E^t_{r,c} + d^t_c - \alpha^t_c(1) - d^t_c - \kappa^t)/\tau\right)}{\sum_{k\in\mathcal{C}_t}\exp\Big((S^t_{r,k} + E^t_{r,k} + d^t_k - \alpha^t_k(1) - d^t_k - \kappa^t)/\tau\Big)}$$

is identical to the predictive assignment induced by $\mathbf{S}^t + \mathbf{E}^t$ and $\boldsymbol{\alpha}^t(1)$, uniqueness of the dual correction modulo a class-independent offset implies

$$\overline{\boldsymbol{\alpha}}^t = \boldsymbol{\alpha}^t(1) + \mathbf{d}^t + \kappa^t \mathbf{1}_{K_t}$$

for some $\kappa^t \in \mathbb{R}$. Since $\boldsymbol{\alpha}^t(0)$ represents the correction associated with $\mathbf{S}^t$, Eq. (10) implies

$$\min_{\kappa^t\in\mathbb{R}} \left\|\overline{\boldsymbol{\alpha}}^t - \boldsymbol{\alpha}^t - \mathbf{d}^t - \kappa^t\mathbf{1}_{K_t}\right\|_2 \le \frac{\sqrt{K_t}}{2\gamma^t}\|\mathbf{E}^t\|_\infty. \tag{11}$$

This establishes that the shared preference component is absorbed exactly by the class-wise correction, while only the sample-specific residual contributes to correction error.

We finally propagate Eq. (11) to calibrated generative evidence. Choose κ^t attaining the minimum in Eq. (11) and define

$$\boldsymbol{\eta}^t = \overline{\boldsymbol{\alpha}}^t - \boldsymbol{\alpha}^t - \mathbf{d}^t - \kappa^t\mathbf{1}_{K_t}.$$

For any $c, c' \in \mathcal{C}_t$, substituting $\overline{\mathbf{s}}^t_i = \mathbf{s}^t_i + \mathbf{d}^t + \mathbf{e}^t_i$ and the definition of $\boldsymbol{\eta}^t$ gives

$$\begin{aligned}\left[(\overline{s}^t_{i,c} - \overline{\alpha}^t_c) - (\overline{s}^t_{i,c'} - \overline{\alpha}^t_{c'})\right] - \left[(s^t_{i,c} - \alpha^t_c) - (s^t_{i,c'} - \alpha^t_{c'})\right] \\ = (e^t_{i,c} - e^t_{i,c'}) - (\eta^t_c - \eta^t_{c'}).\end{aligned} \tag{12}$$

Taking absolute values in Eq. (12), using $\|\boldsymbol{\eta}^t\|_\infty \le \|\boldsymbol{\eta}^t\|_2$, and applying Eq. (11) yield

$$\begin{aligned}&\max_{c,c'\in\mathcal{C}_t}\left|\left[(\overline{s}^t_{i,c} - \overline{\alpha}^t_c) - (\overline{s}^t_{i,c'} - \overline{\alpha}^t_{c'})\right] - \left[(s^t_{i,c} - \alpha^t_c) - (s^t_{i,c'} - \alpha^t_{c'})\right]\right| \\ &\quad\le 2\|\mathbf{e}^t_i\|_\infty + 2\|\boldsymbol{\eta}^t\|_2 \\ &\quad\le 2\|\mathbf{e}^t_i\|_\infty + \frac{\sqrt{K_t}}{\gamma^t}\|\mathbf{E}^t\|_\infty.\end{aligned} \tag{13}$$

Applying $\|\mathbf{E}^t\|_\infty \le \epsilon$ and $\|\mathbf{e}^t_i\|_\infty \le \epsilon$ to Eq. (13) gives

$$\max_{c,c'\in\mathcal{C}_t}\left|\left[(\overline{s}^t_{i,c} - \overline{\alpha}^t_c) - (\overline{s}^t_{i,c'} - \overline{\alpha}^t_{c'})\right] - \left[(s^t_{i,c} - \alpha^t_c) - (s^t_{i,c'} - \alpha^t_{c'})\right]\right| \le \left(2 + \frac{\sqrt{K_t}}{\gamma^t}\right)\epsilon,$$

which is exactly Eq. (4).

When $\epsilon = 0$, both $\mathbf{E}^t$ and $\mathbf{e}^t_i$ vanish. Equation (7) then gives $\overline{\mathbf{Q}}^{t*} = \mathbf{Q}^{t*}$, while Eq. (11) gives $\overline{\boldsymbol{\alpha}}^t = \boldsymbol{\alpha}^t + \mathbf{d}^t + \kappa^t\mathbf{1}_{K_t}$ for some class-independent offset κ^t. Substituting these relations into the calibrated evidence gives $\overline{\mathbf{s}}^t_i - \overline{\boldsymbol{\alpha}}^t = \mathbf{s}^t_i - \boldsymbol{\alpha}^t - \kappa^t\mathbf{1}_{K_t}$, so every pairwise difference in calibrated generative evidence is preserved exactly. Finally, let $\hat{y}^t_i = \arg\max_{c\in\mathcal{C}_t}(s^t_{i,c} - \alpha^t_c)$ and define its calibrated margin as $\Delta^t_i = (s^t_{i,\hat{y}^t_i} - \alpha^t_{\hat{y}^t_i}) - \max_{c\neq\hat{y}^t_i}(s^t_{i,c} - \alpha^t_c)$. Equation (4) implies that every pairwise calibrated difference involving $\hat{y}^t_i$ can decrease by at most $(2 + \sqrt{K_t}/\gamma^t)\epsilon$. Therefore, whenever $\Delta^t_i > (2 + \sqrt{K_t}/\gamma^t)\epsilon$, every perturbed pairwise difference between $\hat{y}^t_i$ and a competing class remains strictly positive, and the predicted class is unchanged. □

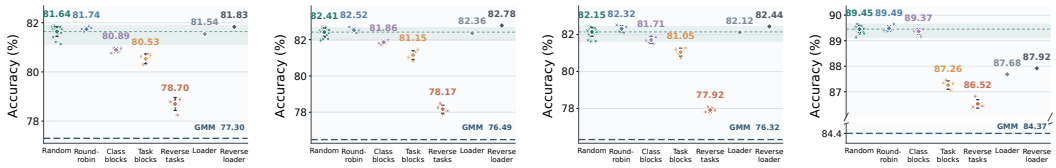
B Implementation Details

Following (Cao et al., 2024), we optimize only the vision-to-language projection while keeping all other components frozen. ImageNet-R is trained for two epochs per task with a learning rate of 3×10^{-7}, while Tiny-ImageNet uses 3×10^{-6} for the initial task and 3×10^{-7} thereafter, with AdamW and cosine decay. Candidate Scoring evaluates all accumulated classes using the template `This is a photo of [CLASS].`, reusing the cached image-question prefix and evaluating candidate answers in chunks of 50. For evaluation after each incremental task, the calibration queue is initialized empty with capacity $M = 800$, and test samples are processed sequentially over the

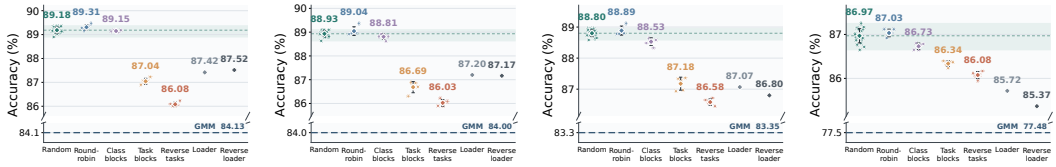
accumulated label space. Online Calibration estimates the class-wise correction from queued generative evidence using Eq. (3), with $\boldsymbol{\alpha}^t = \mathbf{0}$ when the queue is empty, and inserts the current evidence only after prediction. Unless otherwise specified, we use the uniform target marginal $\pi_c^t = 1/K_t$, and Sinkhorn matching is implemented in the log domain with $\tau = 0.3$ across all experiments, a convergence threshold 10^{-6} and at most 200 iterations. All history-based calibration methods use the same generative evidence from previously processed samples and the same evaluation order.

C Further Analysis

Sensitivity to Evaluation Order. Online Calibration estimates persistent generative preference from generative evidence of previously processed samples, making the composition of the calibration queue potentially dependent on the evaluation order. Figure 8 evaluates this sensitivity using random, round-robin, class-block, task-block, reversed-task, loader, and reversed-loader orders across different incremental settings. UnL^2 consistently outperforms GMM under all evaluated orders, indicating that its gains do not depend on a particular evaluation order. Performance variations become more pronounced under highly structured orders, particularly when samples from the same classes are locally concentrated, which reduces the diversity of queued generative evidence available for estimating persistent generative preference.

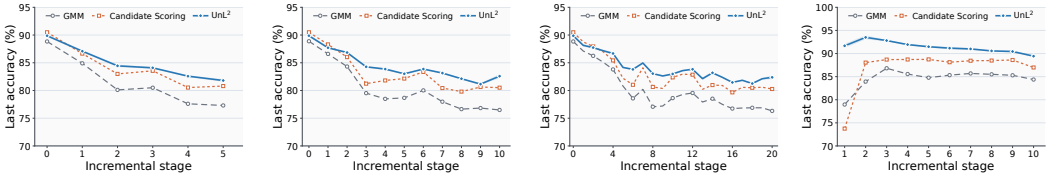


(a) Tiny-ImageNet, 5 tasks (b) Tiny-ImageNet, 10 tasks (c) Tiny-ImageNet, 20 tasks (d) ImageNet-R, 10 tasks

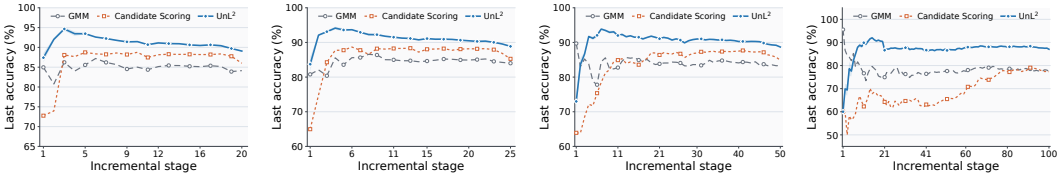


(e) ImageNet-R, 20 tasks (f) ImageNet-R, 25 tasks (g) ImageNet-R, 50 tasks (h) ImageNet-R, 100 tasks

Figure 8: Sensitivity to evaluation order under different numbers of tasks.



(a) Tiny-ImageNet, 5 tasks (b) Tiny-ImageNet, 10 tasks (c) Tiny-ImageNet, 20 tasks (d) ImageNet-R, 10 tasks



(e) ImageNet-R, 20 tasks (f) ImageNet-R, 25 tasks (g) ImageNet-R, 50 tasks (h) ImageNet-R, 100 tasks

Figure 9: Task-wise accuracy under different incremental settings on different datasets.

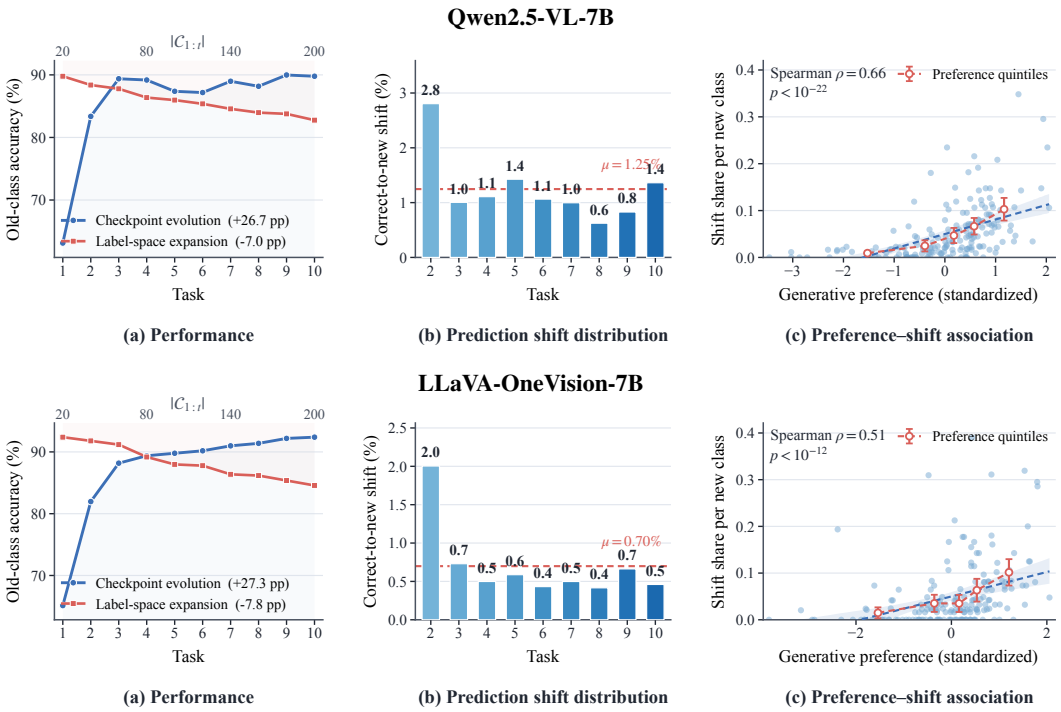


Figure 10: **Generalization of the motivation analysis across MLLM backbones on ImageNet-R.** For both Qwen2.5-VL-7B and LLaVA-OneVision-7B: (a) old-class accuracy improves across successive parameter updates under a fixed label space, but degrades as the candidate label space expands; (b) previously correct old-class predictions shift disproportionately toward newly introduced classes; and (c) prediction shifts are positively associated with class-wise generative preference. These results establish the generality of inference interference across different MLLM backbones.

Algorithm 1 UnL^2 Inference

Require: Model θ^t, accumulated label space $\mathcal{C}_t$, test samples $\{\mathbf{x}_i^t\}$, queue capacity M, target marginal $\boldsymbol{\pi}^t$, entropy coefficient τ
Ensure: Predictions $\{\hat{y}_i^t\}$

Initialize calibration queue $\mathcal{M}^t \leftarrow \emptyset$
for each test sample $\mathbf{x}_i^t$ in evaluation order **do**
  $\mathbf{s}_i^t \leftarrow \mathrm{CandidateScoring}(\mathbf{x}_i^t, \mathcal{C}_t, \theta^t)$
  if $\mathcal{M}^t = \emptyset$ **then**
    $\boldsymbol{\alpha}^t \leftarrow \mathbf{0}$
  else
    Form $\mathbf{S}^t$ from the generative evidence in $\mathcal{M}^t$
    $\boldsymbol{\alpha}^t \leftarrow \mathrm{SinkhornCorrection}(\mathbf{S}^t, \boldsymbol{\pi}^t, \tau)$
  end if
  $\hat{y}_i^t \leftarrow \arg\max_{c \in \mathcal{C}_t} (s_{i,c}^t - \alpha_c^t)$
  Enqueue $\mathbf{s}_i^t$ into $\mathcal{M}^t$
  if $|\mathcal{M}^t| > M$ **then**
    Remove the oldest entry from $\mathcal{M}^t$
  end if
end for
return $\{\hat{y}_i^t\}$