Enhancing Few-Shot Class-Incremental Learning via Training-Free Bi-Level Modality Calibration

Yiyang Chen[1], Tianyu Ding[2], Lei Wang[3], Jing Huo[1], Yang Gao[1], Wenbin Li[1,4*]
[1]State Key Laboratory for Novel Software Technology, Nanjing University, China
[2]Applied Sciences Group, Microsoft, USA [3]University of Wollongong, Australia
[4]Shenzhen Research Institute of Nanjing University, Shenzhen, China

CVPR 2025

Background

- Class-Incremental Learning:

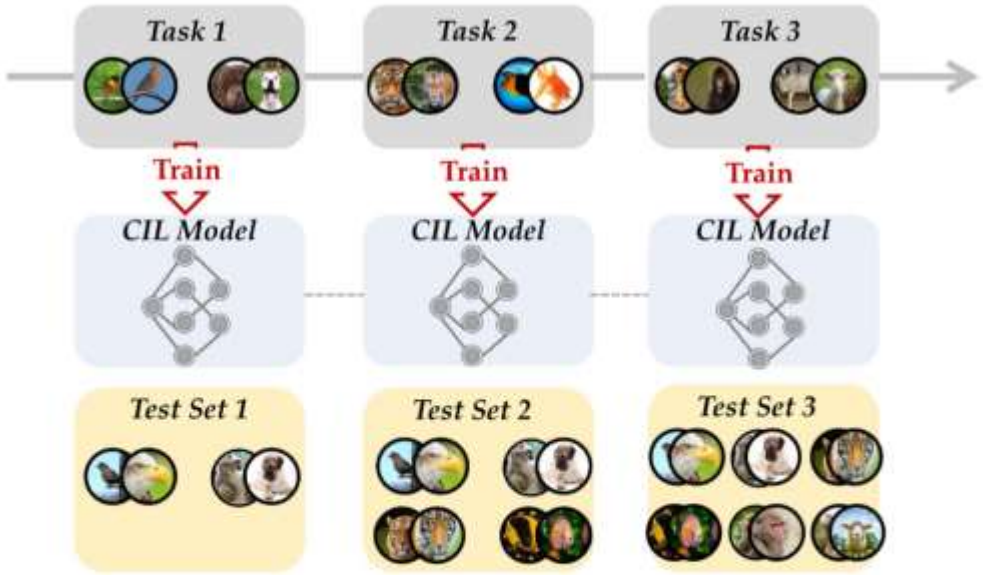


Fig. 1. The setting of CIL. Non-overlapping classes arrive sequentially, and the model needs to learn to classify all the classes incrementally. After learning each task, the model is evaluated among all seen classes. An ideal model should perform well in the newly learned classes and remember the former without forgetting.

Major Challenges:
- catastrophic forgetting
- overfitting

- Few-Shot Class-Incremental Learning (FSCIL):

  Base phase(task 0): sufficient samples
  Incremental phase(task 1~n): a few samples

Motivation

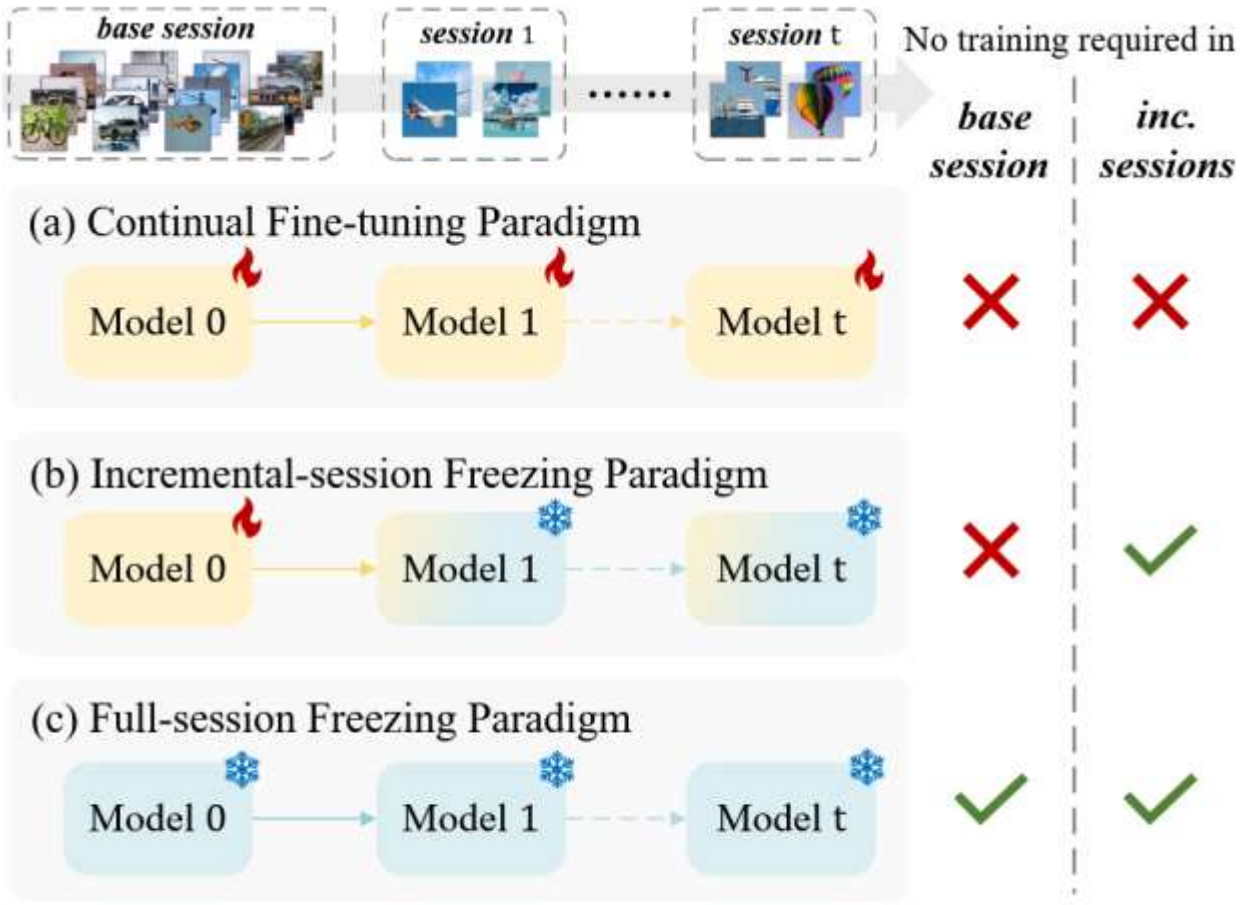
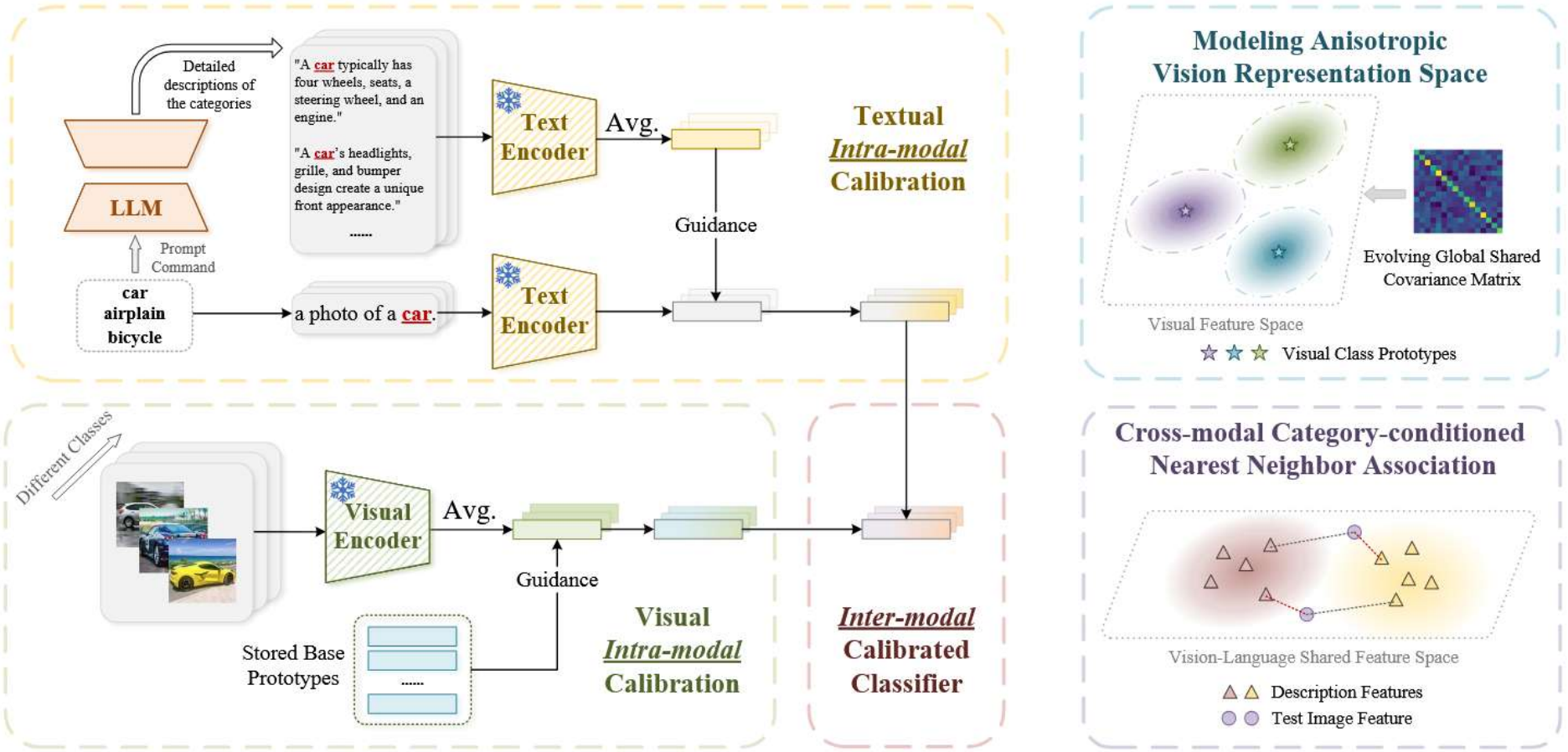


Figure 1. Model updating paradigms in FSCIL. (a) Continuous model updates during incremental phases. (b) Leveraging the base task to learn a generalizable model. (c) Our proposed framework: training-free in both base and incremental sessions.

How can we effectively leverage pre-trained knowledge and domain-specific visual priors to adapt the large-scale model using few samples without additional training?

Bi-level **M**odality **C**alibration (BiMC)

Method



(a) Bi-Level Modality Calibration Framework

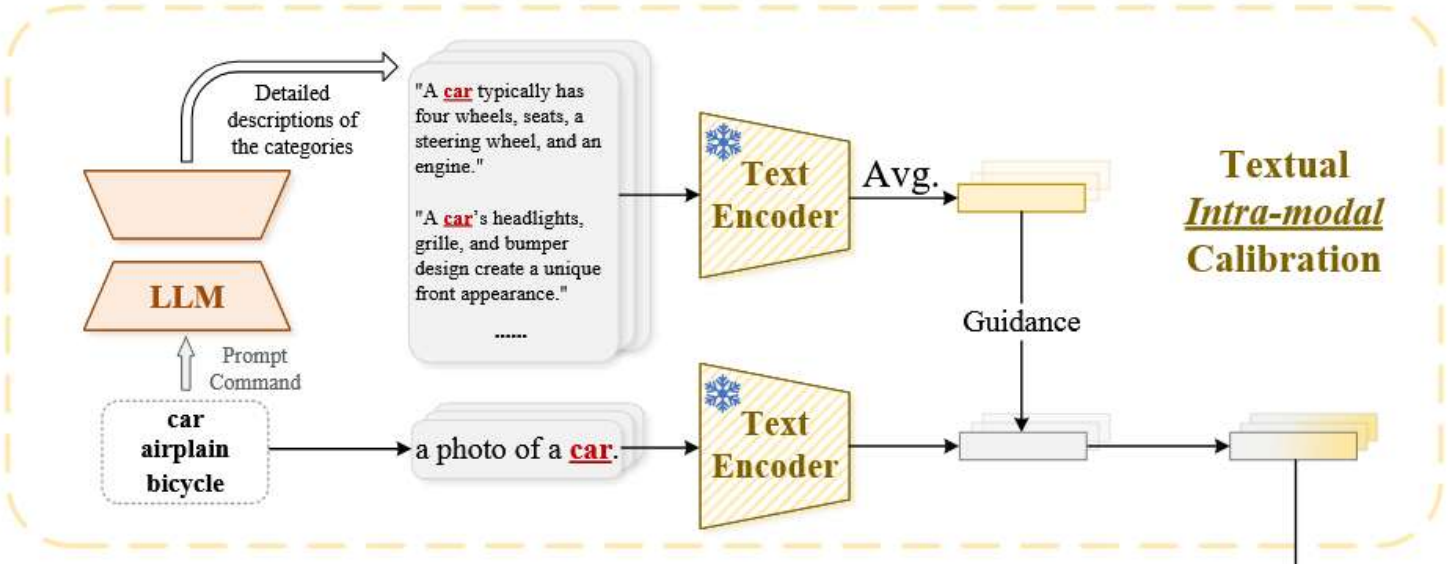
(b) Semantic and Covariance-enhanced Metric

Figure 2. Overview of our framework. In each task adaptation phase, we construct a calibrated classifier through our proposed bi-level calibration framework, which involves *intra-modal calibration* and *inter-modal calibration*. To enhance performance, we introduce a globally shared covariance metric for visual feature modeling, complemented by a category-conditioned nearest-neighbor scoring strategy.

Method

Intra-modal classifier calibration

▶ Textual modality



$$\tilde{\boldsymbol{\mu}}_c^T = (1-\lambda_T)\boldsymbol{w}_c + \lambda_T \left(\frac{1}{n_c} \sum_{j=1}^{n_c} \frac{g\left(\mathbf{t}_{c,j}\right)}{\|g\left(\mathbf{t}_{c,j}\right)\|_2} \right)$$

"a photo of a [CLS]."

These templates fails to capture category-specific nuances, making them particularly inadequate for fine-grained classification tasks

λ_T : the intensity of intra-modal calibration within the text modality

$t_{c,j}$: j-th LLM-generated description for class c

n_c : the number of descriptions for class c

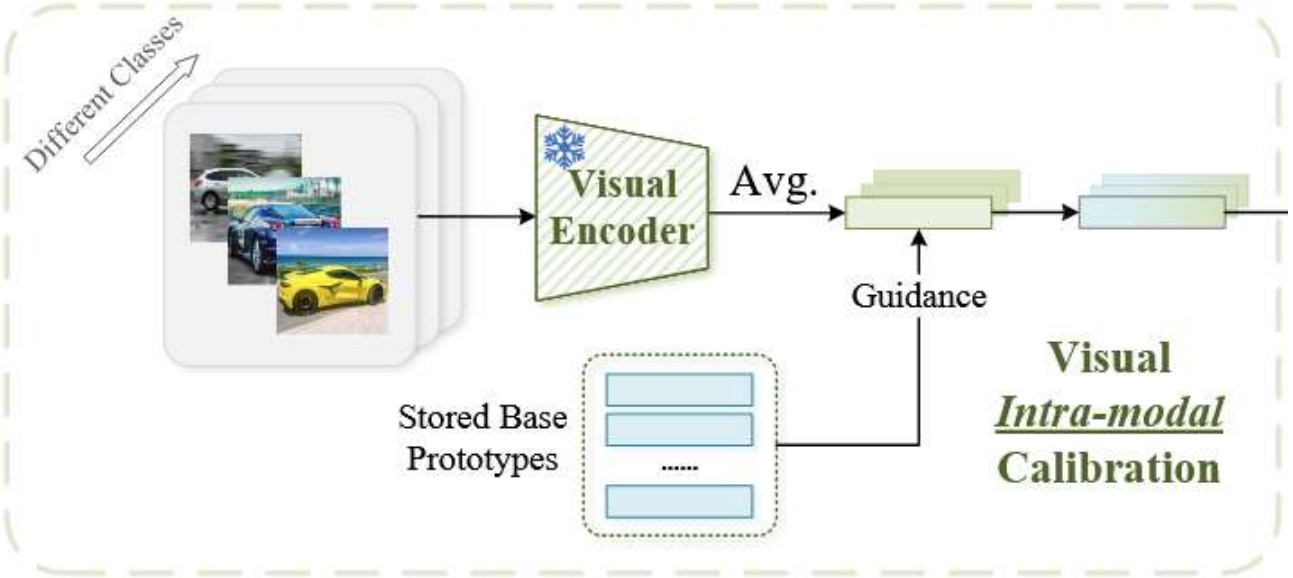
$g(\cdot)$: the text encoder

Method

Intra-modal classifier calibration

the limited number of samples in incremental phase inevitably leads to biased estimates of new class prototypes. → using base class to calibrate

▶ Visual modality



$$\tilde{\boldsymbol{\mu}}_c^I = \begin{cases} \boldsymbol{\mu}_c^I & , t = 0 \\ (1-\lambda_I)\boldsymbol{\mu}_c^I + \lambda_I \sum_{b=1}^{|\mathcal{Y}_0|} s_{b,c}\boldsymbol{\mu}_b^I & , t > 0, \end{cases}$$

λ_I : the strength of the visual intra-modal calibration

$s_{b,c}$: the normalized cosine similarity between the visual prototypes of class b (**from the base classes**) and class c

$<\cdot,\cdot>$: cosine similarity

where

$$\boldsymbol{\mu}_c^I = \frac{1}{m_c}\sum_{j=1}^{m_c}\frac{f(\mathbf{x}_{c,j})}{\|f(\mathbf{x}_{c,j})\|_2}, \qquad s_{b,c} = \frac{e^{\tau\cdot\langle\boldsymbol{\mu}_b,\boldsymbol{\mu}_c\rangle}}{\sum_{i=1}^{|\mathcal{Y}_0|} e^{\tau\cdot\langle\boldsymbol{\mu}_i,\boldsymbol{\mu}_c\rangle}}$$

μ_c^I: naive visual prototype of class c.

m_c : the number of training images in class c

$f(\cdot)$: the visual encoder

$x_{c,j}$: j-th image belonging to class c

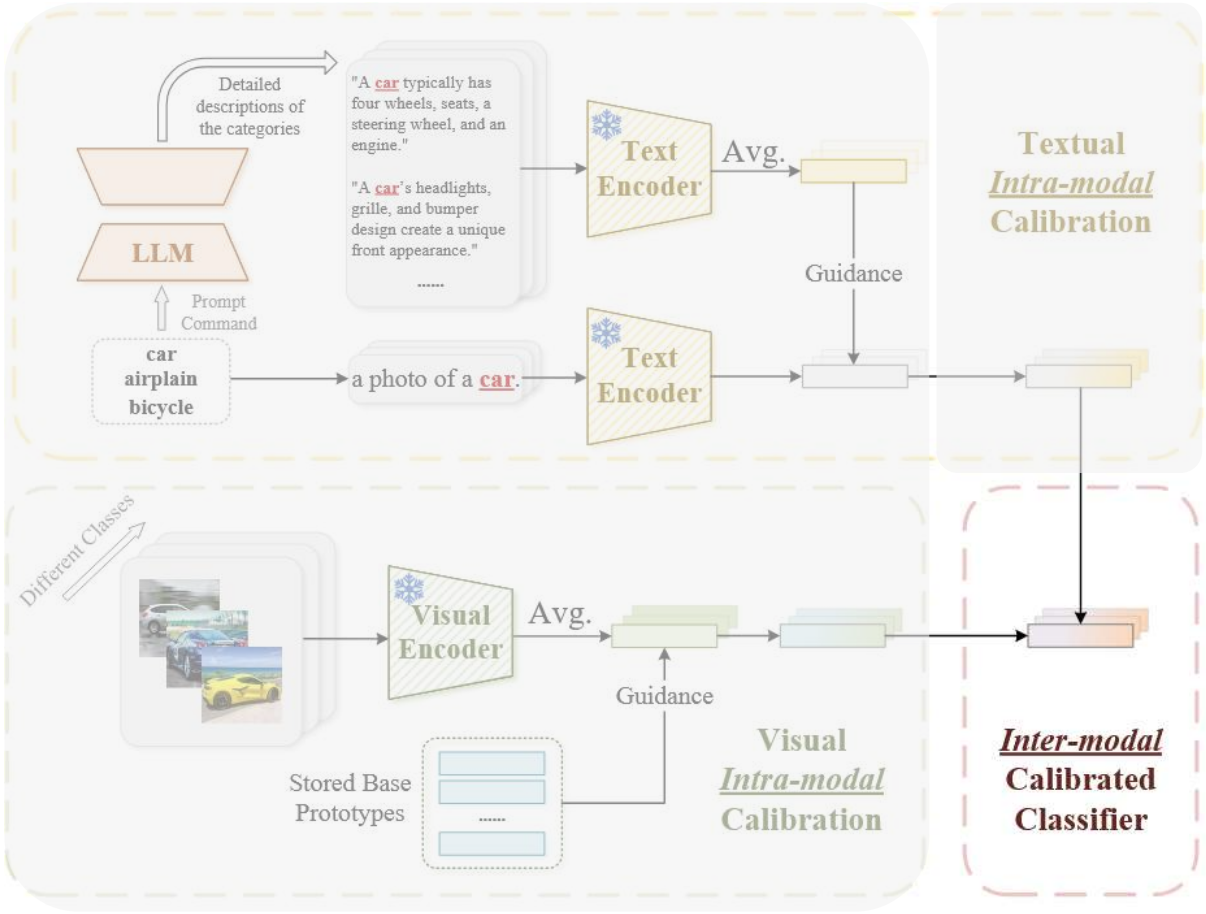
Method

Inter-modal classifier calibration

$$\boldsymbol{\mu}_c = \beta \tilde{\boldsymbol{\mu}}_c^T + (1-\beta)\tilde{\boldsymbol{\mu}}_c^I.$$

- Inference

$$\mathbf{s}_c^{\text{calib}} = \frac{f(\mathbf{x})^\top \boldsymbol{\mu}_c}{\|f(\mathbf{x})\|_2 \cdot \|\boldsymbol{\mu}_c\|_2}.$$



Method

Semantic and covariance-enhanced metric

previous metrics effectively mitigate modal bias, but they primarily measure sample distances from distribution centers. This isotropic approach struggles to capture higher-order information and complex data relationships.

leveraging statistical information from existing data for more comprehensive and efficient measurements

avoid singularity

- For task t:

$$\tilde{\boldsymbol{\Sigma}}^t = \boldsymbol{\Sigma}^t + \frac{\gamma}{d}\,\mathrm{tr}(\boldsymbol{\Sigma}^t)\boldsymbol{I}_d,$$

γ: the strength of regularization

d: feature dimension

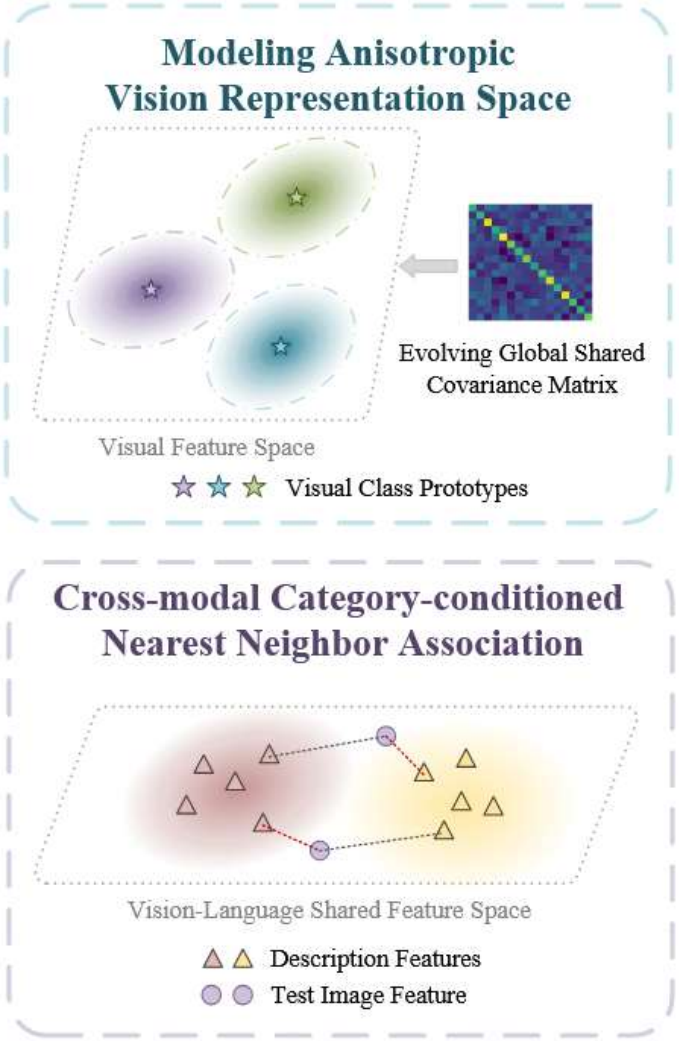
$tr(\cdot)$: trace operator of a matrix

continuously evolving shared covariance matrix

$$\tilde{\boldsymbol{\Sigma}}_G^t = \frac{|\mathcal{Y}_{t-1}|}{|\mathcal{Y}_t|}\tilde{\boldsymbol{\Sigma}}_G^{t-1} + \left(1 - \frac{|\mathcal{Y}_{t-1}|}{|\mathcal{Y}_t|}\right)\tilde{\boldsymbol{\Sigma}}^t.$$

Given a test sample x, calculate its score as:

$$\mathbf{s}_c^{\mathrm{cov}}(\mathbf{x}) = -\frac{1}{d}\left(f(\mathbf{x}) - \tilde{\boldsymbol{\mu}}_c^I\right)^T \tilde{\boldsymbol{\Sigma}}_G^{-1}\left(f(\mathbf{x}) - \tilde{\boldsymbol{\mu}}_c^I\right).$$



(b) Semantic and Covariance-enhanced Metric

Method

Cross-modal category nearest neighbor metric

for new classes

- ✓ To better leverage LLM-generated category descriptions

$$\mathbf{s}_c^{\mathrm{nn}} = \max_j \left\{ \boldsymbol{z}_{c,j}^\top \cdot \boldsymbol{v} \right\},$$

where $\boldsymbol{z}_{c,j} = \frac{g(\mathbf{t}_{c,j})}{\|g(\mathbf{t}_{c,j})\|_2}$ $\boldsymbol{v} = \frac{f(\mathbf{x})}{\|f(\mathbf{x})\|_2}$.

Inference score reorganization strategy

base-derived covariance matrices perform poorly in distinguishing new class data

- covariance from extensive base data fails to align with new classes
- the limited data available for new classes prevents accurate covariance estimation

$$\mathbf{p}_c = \begin{cases} \alpha \mathbf{p}_c^{\mathrm{calib}} + (1-\alpha)\mathbf{p}_c^{\mathrm{cov}} & , c \in \mathcal{Y}_0 \\ \alpha \mathbf{p}_c^{\mathrm{calib}} + (1-\alpha)\mathbf{p}_c^{\mathrm{nn}} & , c \notin \mathcal{Y}_0. \end{cases}$$

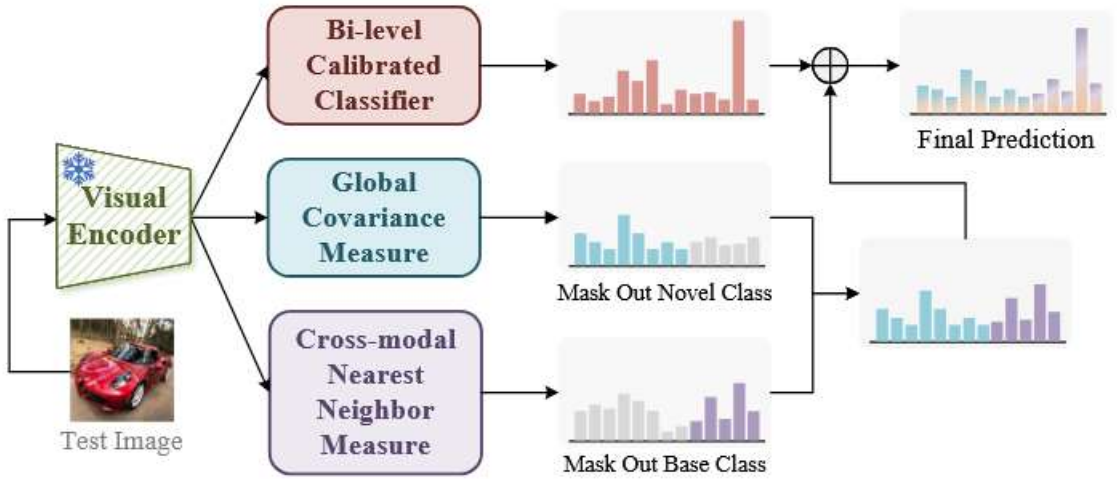


Figure 3. In the inference phase, an ensemble strategy with category masking is utilized.

Experiments

Table 1. Detailed session-wise accuracy, average accuracy (Avg) and performance degradation (PD) comparison on *mini*ImageNet dataset. **V** and **L** represent the visual and language modalities, respectively. BiMC refers to the results obtained solely through the bi-level calibration framework, whereas BiMC† incorporates the ensemble classifier strategy. The best results on each each sessions are indicated in **bold**, while the second-best results are underlined. ↑ means higher is better, while ↓ means lower is better.

Method	Modality	Accuracy in each session(%) ↑									Avg ↑	PD ↓
		0	1	2	3	4	5	6	7	8		
CLIP Zero-Shot [24]	**L**	91.27	91.25	89.74	89.43	88.98	88.42	86.96	86.64	86.15	88.76	5.12
Visual Prototype [29]	**V**	92.58	91.88	89.37	88.24	88.42	87.86	86.33	86.16	86.09	88.55	6.49
TEEN [35]	**V**	92.58	92.06	89.69	88.60	88.80	88.31	86.84	86.71	86.66	88.92	5.92
FeCAM [8]	**V**	94.42	93.71	91.61	90.59	90.62	90.22	88.43	88.17	88.20	90.66	6.22
BiMC	**V-L**	94.90	94.80	93.21	92.91	92.88	92.60	91.78	91.88	91.81	92.97	3.09
BiMC†	**V-L**	**95.47**	**95.34**	**93.80**	**93.63**	**93.51**	**93.18**	**92.51**	**92.56**	**92.40**	**93.60**	**3.07**

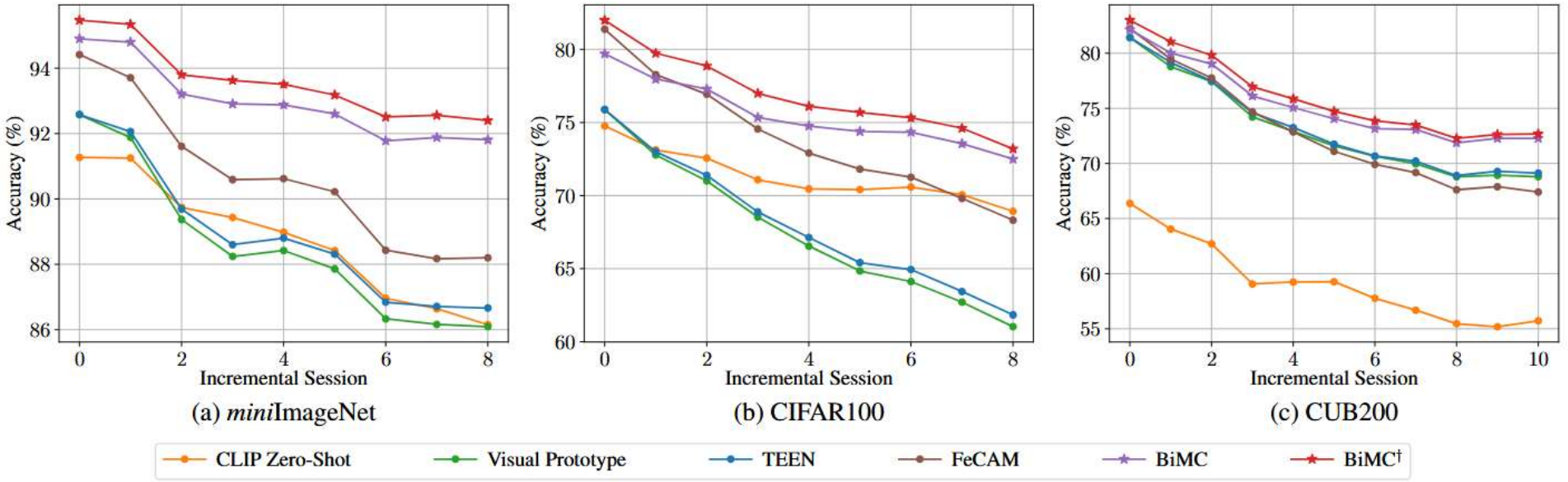


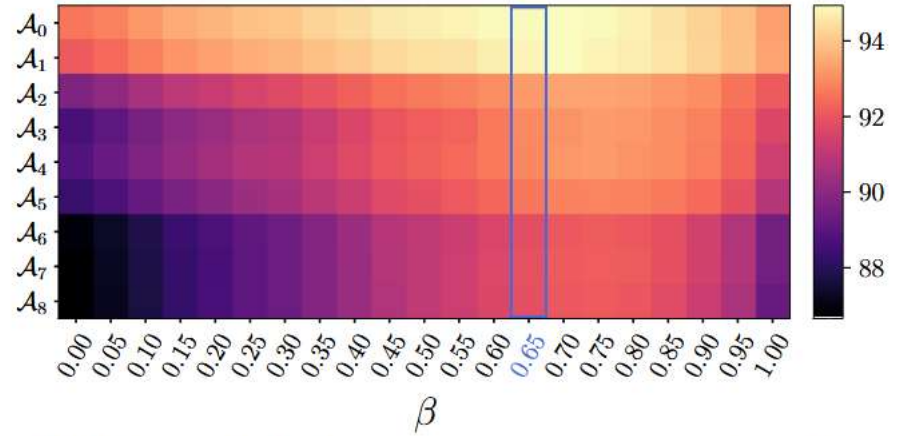
Figure 4. Performance curve of each incremental sessions on (a) *mini*ImageNet, (b) CIFAR100 and (c) CUB200 datasets.

Comparison with trainable methods

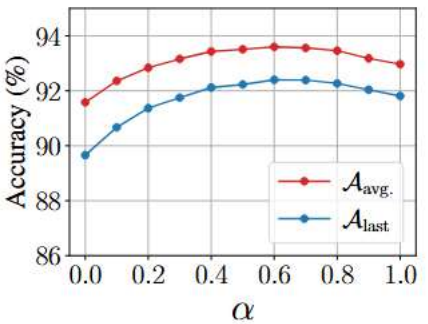
Table 5. Comparison with three trainable methods on the *mini*ImageNet dataset. N_p is the number of parameters which require training. $\Delta\mathcal{A}_{\text{last}}$ reflects the last task's performance gap between our method and the comparative one.

Method	N_p	$\mathcal{A}_{\text{base}}$	$\mathcal{A}_{\text{last}}$	$\mathcal{A}_{\text{avg.}}$	$\Delta\mathcal{A}_{\text{last}}$
CPE-CLIP [5]	400k	90.23	82.77	86.13	+9.63
CLIP-M^3 [6]	46k	96.00	92.50	94.10	−0.10
LP-DiF [10]	8.1k	96.34	91.68	93.76	+0.72
BiMC†	**0**	95.47	92.40	93.60	0.00

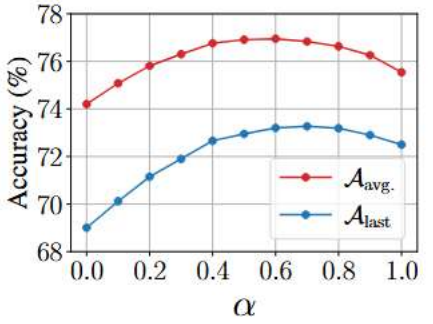
Analysis on hyper-parameters



(a) Influence of β and session-wise accuracy on *mini*ImageNet



(b) Influence of α on *mini*ImageNet



(c) Influence of α on CIFAR100

Figure 7. Analysis on hyper-parameters β and α. For β, we present heatmaps illustrating accuracy under varying β across different sessions. For α, we report the average task accuracy $\mathcal{A}_{avg.}$ and the final task accuracy $\mathcal{A}_{last}$ under different settings.

Ablation study

Table 2. Ablation studies of bi-level calibration framework on *mini*ImageNet. **Vis.** and **Lang.** individually represent classifiers derived from the visual and language modalities. **Intra-C** and **Inter-C** refer to the strategies of intra-modal and inter-modal calibration, respectively.

Vis	Lang	Intra-C	Inter-C	Accuracy in each session(%) ↑									Avg ↑	PD ↓
				0	1	2	3	4	5	6	7	8		
✓				92.58	91.88	89.37	88.24	88.42	87.86	86.33	86.16	86.09	88.55	6.49
✓		✓		92.58	92.06	89.69	88.60	88.80	88.31	86.84	86.71	86.66	88.92	5.92
	✓			91.27	91.25	89.74	89.43	88.98	88.42	86.96	86.64	86.15	88.76	5.12
	✓	✓		93.27	93.32	91.91	91.51	91.21	90.73	89.28	89.23	89.01	91.05	4.26
✓	✓		✓	94.68	94.58	92.87	92.49	92.56	92.31	91.59	91.58	91.54	92.69	3.14
✓	✓	✓	✓	**94.90**	**94.80**	**93.21**	**92.91**	**92.88**	**92.60**	**91.78**	**91.88**	**91.81**	**92.97**	**3.09**

Table 4. Ablation study of Semantic and Covariance-enhanced Metric on the *mini*ImageNet dataset. To verify their efftctiveness, we measured the performance of the base task $\mathcal{A}_{\text{base}}$, the performance of last task $\mathcal{A}_{\text{last}}$ and the average performance $\mathcal{A}_{\text{avg.}}$ across all tasks. Additionally, we report the accuracy of base $\mathcal{A}^{b}_{\text{last}}$ and novel $\mathcal{A}^{n}_{\text{last}}$ class in last task.

Ablation	$\mathcal{A}_{\text{base}}$	$\mathcal{A}_{\text{last}}$	$\mathcal{A}^{b}_{\text{last}}$	$\mathcal{A}^{n}_{\text{last}}$	$\mathcal{A}_{\text{avg.}}$
BiMC	94.90	91.81	93.43	89.38	92.97
+ MGC	**95.47**	91.51	**94.37**	87.22	92.91
+ CMNN	94.85	92.15	93.30	90.42	93.18
BiMC† (w/o mask.)	**95.47**	92.20	94.12	89.32	93.37
BiMC†	**95.47**	**92.40**	93.20	**91.20**	**93.60**

Thanks!