BEYOND PRIMITIVE REUSE: PRIMITIVE DISCOVERY AND ADAPTIVE MODULATION FOR FEW-SHOT CLASS-INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot class-incremental learning (FSCIL) is essential for adapting to open-world environments. Humans can rapidly learn novel concepts by decomposing objects into key components and recombining familiar components through associative memory. Inspired by this ability, compositional approaches reuse base-class primitives or update the decomposition module in subsequent sessions. However, whole-primitive reuse may overlook discriminative details, while jointly learning decomposition and classification is challenging with few examples. To address this limitation, we propose Primitive Discovery and Adaptive Modulation (**PDAM**) for FSCIL. Specifically, Object-centric Primitive Discovery (OPD) module learns to decompose objects into primitives from base-class images and transfers this capability to novel categories. However, the decomposition alone does not ensure effective recognition, as the resulting primitives contain both shared appearance and potentially discriminative details whose relevance is difficult to establish from few examples. We therefore introduce a Primitive-guided Adaptive Modulation (PAM) module that adaptively modulates features across regions and channels to support discrimination. Experiments demonstrate that our approach achieves state-of-the-art performance on three fine-grained FSCIL benchmarks.

1 INTRODUCTION

Deep neural networks (DNNs) have achieved remarkable progress in visual recognition, yet their success often relies on abundant labeled data and a fixed set of training categories. In open-world environments, new concepts continually emerge (Wang et al., 2020; Yuan et al., 2025), while collecting sufficient labeled examples for each is often impractical. Few-shot class-incremental learning (FSCIL) (Tao et al., 2020) addresses these challenges by learning new categories from limited examples while retaining previously acquired knowledge. Under this setting, models are initially trained on abundant labeled data from base classes and subsequently learn novel classes through incremental sessions, with only a few labeled examples available per new class.

FSCIL faces the dual challenges of *overfitting* to limited novel class examples and *catastrophic forgetting* of previously learned classes. From a representation-learning perspective, existing approaches address these challenges through backbone adaptation (Shi et al., 2021), meta-learning (Chi et al., 2022), prototype calibration (Wang et al., 2026). These approaches primarily operate on image-level representations without explicitly modeling the visual components underlying category recognition. Unlike machines, humans can learn new concepts quickly by decomposing objects into a few key components (Chen et al., 2024) and drawing on hippocampal associative memory (Chandra et al., 2025; Gutiérrez et al., 2024) to recombine previously learned components into new representations. Motivated by human learning, several studies have explored decomposition and recomposition strategies. For instance, Comp-FSCIL (Zou et al., 2024) represents categories as compositions of object parts (i.e., **primitives**) and encourages their reuse across classes to facilitate few-shot recognition in subsequent incremental sessions. In the broader continual learning (Biesialska et al., 2020), CompSLOT (Liao et al., 2026) integrates Slot Attention into models, aggregates the extracted slots into image-level representations, and aligns the relationships with outputs.

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.

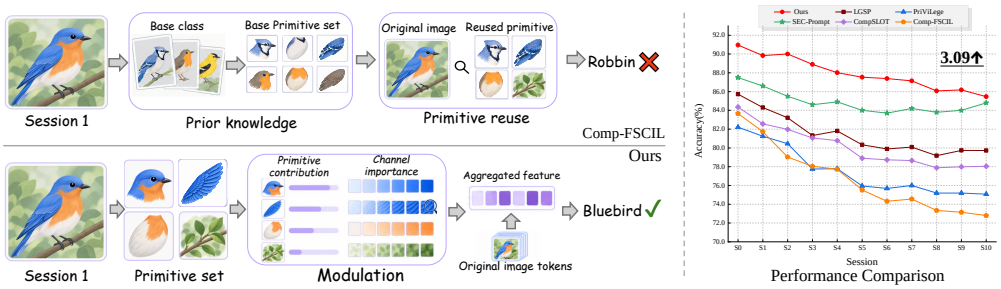


Figure 1: Motivation of the proposed PDAM. **Left:** Composition-based FSCIL reuses primitives from base classes to represent novel images, which may underemphasize distinctive visual details. In contrast, PDAM discovers primitives directly from the current image and adaptively modulates their regional contributions and channel responses. **Right:** Session-wise accuracy on CUB-200 (S0–S10), with a 3.09% average gain over the best baseline.

Despite their effectiveness, existing compositional approaches still face limitations in FSCIL. Comp-FSCIL alleviates data scarcity by reusing primitives across classes, but component-level reuse may overlook discriminative features within each primitive, as illustrated in Fig. 1. This limitation is particularly important in fine-grained recognition. For example, head primitives extracted from "White-crowned Sparrows" and "White-throated Sparrows" may capture similar overall appearances but contain distinct facial markings. Reusing these primitives across classes may obscure such discriminative details when shared appearance dominates their representations, making the two categories harder to distinguish. CompSLOT, by contrast, continues to update its slot decomposition and selection modules in each new task. Under the constraints of FSCIL, learning these modules from only a few novel-class examples may lead to unreliable decomposition and selection.

These limitations motivate us to rethink: *How can we transfer the ability to discover primitives to novel classes while emphasizing discriminative details within them?* To this end, we propose **Primitive Discovery and Adaptive Modulation** (**PDAM**). Inspired by the compositional generalization of Slot Attention (Locatello et al., 2020), we introduce an Object-centric Primitive Discovery (OPD) module that learns to discover visual primitives from base-class images. Once learned, this discovery capability generalizes to novel categories, allowing the frozen OPD module to decompose their images into input-specific primitives without further training (Fig. 3(a)). Beyond discovering primitives, the model must learn how to use them to distinguish novel classes with limited supervision. We therefore introduce Primitive-guided Adaptive Modulation (PAM), which guides this adaptation by weighting primitives and modulating their associated feature channels. The modulated patch features emphasize discriminative cues for learning novel classes. During incremental sessions, we jointly optimize only PAM, visual prompts (Jia et al., 2022), and the classifier, while keeping all remaining parameters frozen to mitigate catastrophic forgetting.

Our main contributions can be summarized as follows: (1) We propose PDAM, a FSCIL framework that transfers the ability to discover primitives from base classes to novel categories and uses the resulting primitives to guide selective feature adaptation under limited supervision. (2) We introduce Primitive-guided Adaptive Modulation (PAM), which combines primitive-level weighting with primitive-conditioned channel modulation to emphasize informative regions and discriminative details for few-shot recognition. (3) Extensive experiments on three fine-grained FSCIL benchmarks demonstrate the state-of-the-art performance of PDAM, with ablation studies and qualitative analyses further supporting the effectiveness of its key components.

2 Related Work

2.1 Few-shot Class-incremental Learning

Few-shot class-incremental learning (FSCIL) requires models to recognize novel categories from limited examples while retaining previously learned knowledge (Tao et al., 2020; Tang et al., 2024; Wang et al., 2026). A major line of research improves the feature space to accommodate future

classes. FACT (Zhou et al., 2022) reserves embedding space through virtual prototypes, while NC-FSCIL (Yang et al., 2023a) maintains feature–classifier alignment using a fixed simplex equiangular tight frame. Another line exploits relationships between base and novel classes to support knowledge transfer. TEEN (Wang et al., 2023) calibrates novel-class prototypes using similarity-weighted base-class prototypes, while Comp-FSCIL (Zou et al., 2024) models categories as compositions of reusable visual primitives. These approaches illustrate the value of base-class knowledge for improving novel-class representations under scarce supervision. With the adoption of pretrained backbones (Han et al., 2022), recent prompt-based methods (Liu et al., 2024; Liu & Yang, 2025; Jiang et al., 2025) adapt representations through learnable prompts to support novel-class recognition while mitigating forgetting of previously learned classes.

2.2 Object-Centric Learning

Object-centric learning (OCL) aims to represent visual scenes as collections of distinct entities, providing a structured basis for compositional understanding. Slot Attention (Locatello et al., 2020) introduces an iterative competitive attention mechanism that groups visual features into exchangeable slots. Subsequent studies extend this approach to increasingly complex visual data. STEVE (Singh et al., 2022) combines slot representations with a Transformer-based decoder to learn object-centric representations from complex videos. For real-world images, DINOSAUR (Seitzer et al., 2022) demonstrates that reconstructing self-supervised pretrained features provides an effective learning signal for discovering objects without region annotations. Building on this direction, SPOT (Kakogeorgiou et al., 2024) improves object grouping through attention-based self-training and patch-order permutation in autoregressive decoding. More recently, SlotAdapt (Akan & Yemez, 2025) incorporates slot-conditioned adapters into pretrained diffusion models to support object discovery and compositional generation. These advances establish effective mechanisms for organizing visual content into structured representations.

3 Methodology

As illustrated in Fig. 2, PDAM comprises two main modules. Object-centric Primitive Discovery (OPD, Sec. 3.1) learns to decompose images into visual primitives from base-class data and transfers this capability to novel categories. Based on the resulting decomposition, Primitive-guided Adaptive Modulation (PAM, Sec. 3.2) weights primitives and modulates their associated feature channels to emphasize discriminative visual cues for recognition.

Problem setup. FSCIL comprises $T+1$ sequential learning sessions, including a base session followed by T incremental sessions. At session $t \in \{0, \ldots, T\}$, the model receives a labeled training set $\mathcal{D}^t_{\text{train}} = \{(\boldsymbol{x}^t_i, y^t_i)\}_{i=1}^{n_t}$, where $\boldsymbol{x}^t_i$ denotes an input image and $y^t_i \in \mathcal{C}^t$ its class label. The class sets are mutually disjoint across sessions, satisfying $\mathcal{C}^t \cap \mathcal{C}^{t'} = \varnothing$ for all $t \neq t'$. The base session ($t = 0$) provides abundant labeled examples for the base classes. Each incremental session ($t \geq 1$) follows an N-way K-shot protocol (e.g., 5-way 5-shot), introducing N novel classes with K labeled examples per class, yielding $|\mathcal{C}^t| = N$ and $n_t = NK$. During session t, only $\mathcal{D}^t_{\text{train}}$ is accessible for training, and training images from previous sessions cannot be revisited. After each session, the model is evaluated on held-out test examples from the cumulative class set $\mathcal{C}^{\leq t} = \bigcup_{s=0}^{t} \mathcal{C}^s$. Each test example must be classified over all classes encountered so far without access to its session identity. The objective is to learn novel classes from limited supervision while preserving recognition performance on previously learned classes.

3.1 Object-Centric Primitives Discovery

Building on Slot Attention (Locatello et al., 2020), we discover primitives from dense patch representations in a self-supervised manner. Through iterative attention, slots compete to aggregate information from different spatial regions, progressively forming a structured representation of the image. This decomposition is learned through competitive reconstruction of pretrained visual representations, without requiring explicit region annotations. Formally, given an image $\boldsymbol{x}$, we first extract patch features using a pre-trained image encoder $f(\cdot)$, yielding $\mathbf{z} = f(\boldsymbol{x}) \in \mathbb{R}^{N_p \times D}$, where N_p denotes the number of embeddings and D is the output embedding dimension. We omit the sample and session indices for clarity.

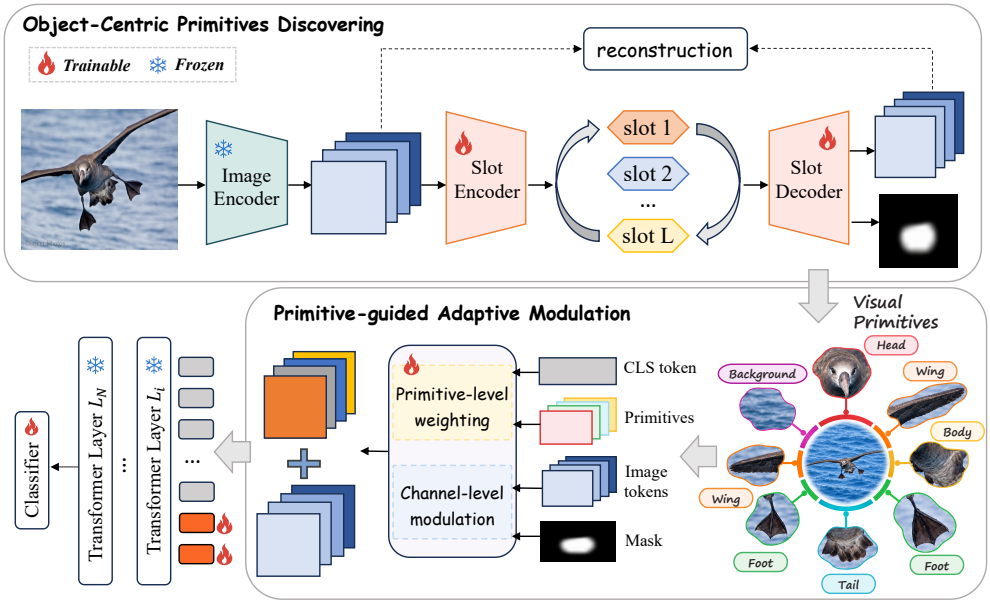


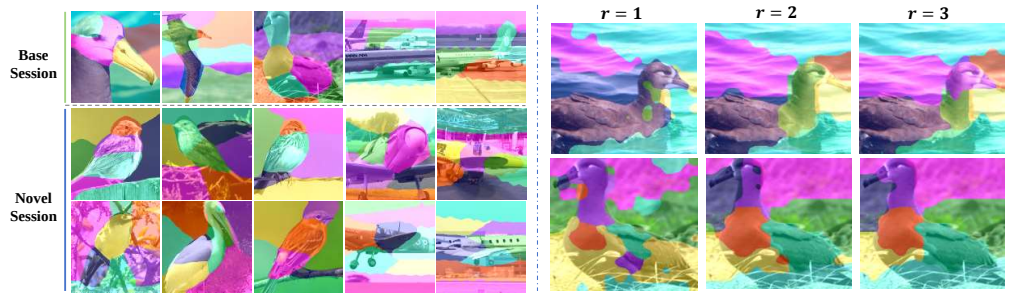
Figure 2: Overview of the proposed PDAM: (1) Object-Centric Primitive Discovery iteratively refines slots into visual primitives through self-supervised feature reconstruction. (2) Primitive-guided Adaptive Modulation combines primitive-level weighting with channel-level modulation to emphasize informative regions and features. The modulated image tokens are then processed by VPT-Deep.

Iterative primitive discovery. We initialize L learnable slots $\mathbf{s} \in \mathbb{R}^{L \times D_s}$ from a standard Gaussian distribution (Goodman, 1963) $\mathcal{N}(\boldsymbol{s}; \boldsymbol{\mu}, \boldsymbol{\sigma})$, where D_s denotes the slot embedding dimension. We then introduce three learnable projection networks $\mathcal{K}_\beta$, $\mathcal{Q}_\gamma$, and $\mathcal{V}_\phi$ to map patch and slot representations into a shared D_s-dimensional attention space. Specifically, the patch representations $\mathbf{z}$ are projected into keys $\mathbf{k} = \mathcal{K}_\beta(\mathbf{z}) \in \mathbb{R}^{N_p \times D_s}$ and values $\mathbf{v} = \mathcal{V}_\phi(\mathbf{z}) \in \mathbb{R}^{N_p \times D_s}$, while the slots are projected into queries $\mathbf{q} = \mathcal{Q}_\gamma(\mathbf{s}) \in \mathbb{R}^{L \times D_s}$. The attention function $\mathrm{attn}(\cdot)$ and the slot update aggregation function $\mathrm{update}(\cdot)$ are defined as follows,

$$\mathrm{update}(\boldsymbol{A}, \mathbf{v}) = \boldsymbol{A}^\top \mathbf{v}, A_{i,j} = \frac{\mathrm{attn}(\mathbf{q}, \mathbf{k})_{ij}}{\sum_{\ell=1}^{N_p} \mathrm{attn}(\mathbf{q}, \mathbf{k})_{\ell j}}, \mathrm{attn}(\mathbf{q}, \mathbf{k}) = \frac{e^{M_{ij}}}{\sum_{\ell=1}^{L} e^{M_{i\ell}}}, \boldsymbol{M} = \frac{\mathbf{k}\mathbf{q}^\top}{\sqrt{D_s}}, \quad (1)$$

where $\boldsymbol{A} \in \mathbb{R}^{N_p \times L}$ is the attention matrix. A gated recurrent unit (GRU) (Dey & Salem, 2017) iteratively updates the slot states by integrating the aggregated patch representations while retaining relevant context from previous iterations. Starting from the initial slots $\mathbf{s}^0$, at each iteration $r \in \{0, \ldots, R-1\}$, we recompute the queries from the slots as $\mathbf{q}^r = \mathcal{Q}_\gamma(\mathbf{s}^r)$ and the slots are then updated as $\mathbf{s}^{r+1} := \mathrm{update}(\mathrm{attn}(\mathbf{q}^r, \mathbf{k}), \mathbf{v})$. After R iterations, the initially coarse slots are progressively refined into meaningful visual primitives, as shown in Fig. 3 (b).

Self-supervised primitive learning. To establish semantic correspondence between the slot representations $\mathbf{s}$ and $\mathbf{z}$, we learn the primitive decomposition in an unsupervised manner through feature reconstruction. After R refinement iterations, the updated slots $\mathbf{s}^R$ are fed into a lightweight MLP-based spatial broadcast decoder $\Phi_d(\cdot)$. The decoder spatially broadcasts each refined slot to the N_P positions and incorporates positional embeddings, yielding $\Phi_d(\mathbf{s}_l^R) \in \mathbb{R}^{N_P \times (D+1)}$. Splitting this output along the channel dimension gives the reconstructed feature embeddings $\hat{\mathbf{z}}_l \in \mathbb{R}^{N_p \times D}$ and the corresponding activation region mask $\alpha_l \in \mathbb{R}^{N_p \times 1}$. To associate each primitive with its corresponding visual region, we apply softmax to α_l across slots at each spatial position, yielding spatial masks $\mathbf{m}_l$. This normalization introduces competition among primitives for each spatial position, encouraging each primitive to associate with a semantically coherent region. The reconstructed fea-



(a) Transferability of Primitive Discovery (b) Iterative Primitive Refinement

Figure 3: Visualization of object-centric primitive discovery. (a) OPD trained on base-class images captures meaningful visual components in both base and novel classes without further training. (b) Primitive assignments become more spatially coherent over successive refinement iterations.

tures are then obtained by combining the primitive-specific predictions according to spatial masks:

$$\hat{\mathbf{z}} = \sum_{l=1}^{L} \hat{\mathbf{z}}_l \odot \mathbf{m}_l, \qquad \mathbf{m}_l = \mathrm{softmax}_l(\alpha_l), \tag{2}$$

where $\odot$ denotes element-wise multiplication with each mask broadcast along the feature dimension. To train the primitive decomposition module and decoder, we minimize the reconstruction error between the reconstructed features $\hat{\mathbf{z}}$ and the target features $\mathbf{z}$:

$$\mathcal{L}_{\mathrm{OPD}} = \|\hat{\mathbf{z}} - \mathbf{z}\|_2^2, \qquad \hat{\mathbf{z}} = \Phi_d(\mathbf{s}) \tag{3}$$

This objective encourages the primitives to collectively preserve the image's visual information while learning their spatial associations without explicit region annotations.

Remark 1 (Semantic feature quality underpins primitive discovery.) *The semantic information in* $\mathbf{z}$ *provides a more effective signal for separating slots into meaningful visual components than reconstructing RGB pixels (Seitzer et al., 2022). We therefore use* $\mathbf{z}$ *as the reconstruction target. The quality of* $\mathbf{z}$ *is consequently an important foundation for effective primitive discovery. Self-supervised pretrained encoders (e.g., DINOv1, DINOv2) (Caron et al., 2021; Oquab et al., 2023) provide transferable patch representations with coherent regional semantics, facilitating the grouping of semantically related patches into slots. An encoder trained with cross-entropy may prioritize class discrimination without preserving clear semantic distinctions between visual regions. This can make it harder for slots to separate distinct visual components.*

3.2 Primitive-Guided Adaptive Modulation

The primitives learned in Sec. 3.1 capture diverse visual components, but their contribution to image reconstruction does not necessarily reflect their relevance to recognition (Tang et al., 2025b;a). For example, background regions may be reconstructed accurately while providing little evidence for distinguishing fine-grained categories. This issue is amplified in novel sessions, where limited supervision makes it difficult to identify subtle discriminative cues. We therefore introduce Primitive-Guided Adaptive Modulation (PAM), which weights different primitives and modulates their associated feature channels to emphasize informative visual evidence.

Primitive-level weighting. We first estimate how strongly each primitive should contribute to feature modulation. Since primitives are unordered, the same primitive index does not necessarily represent a consistent visual component across images. Fixed index-specific weights are therefore unsuitable. Instead, we determine each primitive's contribution from its visual content by relating its representation to the global image context. For each primitive $\mathbf{s}_l$, we obtain its corresponding visual primitive by applying the decoded spatial mask $\mathbf{m}_l$ to the patch features $\mathbf{c}_l = \mathbf{z}_l \odot \mathbf{m}_l \in \mathbb{R}^{N_p \times D}$. Each visual primitive $\mathbf{c}_l$ captures the region-specific visual content associated with $\mathbf{s}_l$

while preserving the spatial structure of the patch features. We aggregate each visual primitive into a local descriptor $\bar{\mathbf{c}}_l = \frac{\sum_{n=1}^{N_p} \mathbf{c}_{ln}}{\sum_{n=1}^{N_p} \mathbf{m}_{ln}+\epsilon} \in \mathbb{R}^D$. To determine the relative contribution of each primitive, we relate its local representation to the global image context. Let $\mathbf{z}_{\text{cls}} \in \mathbb{R}^D$ denote the encoder's CLS token. We compute the context-dependent primitive weights as:

$$e_l = \frac{(\mathbf{W}_g \,\text{sg}(\mathbf{z}_{\text{cls}}))^\top (\mathbf{W}_u \bar{\mathbf{c}}_l)}{\sqrt{d_r}}, \qquad w_l = \frac{\exp(e_l/\tau)}{\sum_{j=1}^{L} \exp(e_j/\tau)}, \tag{4}$$

where $\mathbf{W}_g, \mathbf{W}_u \in \mathbb{R}^{d_r \times D}$ are learnable projections, τ is a temperature parameter, d_r denotes the dimensionality of the projected space, and sg denotes stop-gradient. The resulting weights allocate the modulation contribution among primitives according to their compatibility with the global context. Since the weights are computed from the current image representations, this allocation depends on primitive content rather than a predefined semantic role for each slot.

Channel-level modulation. Primitive weighting determines the relative contribution of each primitive to feature modulation. A region relevant to recognition may contain both features shared across classes and subtle variations that distinguish them. This distinction is particularly important in FSCIL, where the limited labeled examples in novel sessions make it difficult to learn which local features to rely on. We therefore introduce primitive-specific channel modulation to complement primitive weighting with finer-grained feature adjustments. Specifically, the spatial response of channel d is $\mathbf{a}_d = \mathbf{z}_{:,d} \in \mathbb{R}^{N_p}$. We center and normalize both the channel responses and the primitive masks. For a spatial vector $\mathbf{v} \in \{\mathbf{a}_d, \mathbf{m}_l\}$, we define:

$$\mu(\mathbf{v}) = \frac{1}{N_p} \sum_{n=1}^{N_p} v_n, \quad \mathcal{N}(\mathbf{v}) = \frac{\mathbf{v} - \mu(\mathbf{v})\mathbf{1}}{\|\mathbf{v} - \mu(\mathbf{v})\mathbf{1}\|_2 + \epsilon}, \tag{5}$$

where $\mathbf{1} \in \mathbb{R}^{N_p}$ denotes the all-ones vector. Centering removes the spatially uniform component, while normalization reduces the influence of response magnitude, allowing the comparison to focus on spatial variation. We then define a relation descriptor for each primitive: $r_{ld} = \|\mathcal{N}(\mathbf{a}_d) - \mathcal{N}(\mathbf{m}_l)\|_2^2, \quad \mathbf{r}_l = [r_{l1}, \ldots, r_{lD}]^\top \in \mathbb{R}^D$. For non-degenerate responses, a smaller distance indicates closer agreement between the channel's spatial response pattern and the primitive's spatial membership pattern. Thus, $\mathbf{r}_l$ describes how different channels relate spatially to primitive l. Spatial correspondence alone does not establish discriminative importance. We therefore use a shared learnable mapping to convert the relation descriptor into signed channel adjustments $\mathbf{g}_l = \tanh(f_\theta(\mathbf{r}_l)) \in (-1, 1)^D$, where f_θ is a two-layer MLP shared across primitives. This mapping learns how the observed spatial relations should influence channel responses, allowing both enhancement and attenuation of individual feature channels.

Finally, we combine the primitive weights, spatial masks, and channel adjustments into a unified modulation matrix. We formulate the modulation as follows:

$$\mathbf{G} = \mathbf{M}^\top \,\text{Diag}(\mathbf{w})\mathbf{Q} \in \mathbb{R}^{N_p \times D}, \qquad \widetilde{\mathbf{z}} = \mathbf{z} + \lambda(\mathbf{G} \odot \mathbf{z}), \tag{6}$$

where $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_L]^\top \in \mathbb{R}^{L \times N_p}$, $\mathbf{Q} = [\mathbf{g}_1, \ldots, \mathbf{g}_L]^\top \in \mathbb{R}^{L \times D}$, $\mathbf{w} = [w_1, \ldots, w_L]^\top \in \mathbb{R}^L$, $\text{Diag}(\mathbf{w})$ denotes the diagonal matrix of primitive weights. The modulation strength λ is a learnable scalar jointly optimized with the other PAM parameters. This formulation combines the spatial support of each primitive with its channel adjustments, weighted by its relative contribution.

3.3 Prompt-Based Adaptation and Optimization

We feed the modulated patch features $\widetilde{\mathbf{z}}$ together with the CLS token into subsequent Transformer blocks. Following VPT-Deep (Jia et al., 2022), we insert learnable prompt tokens $\mathcal{P}$ at selected later blocks to obtain the image representation $\mathbf{v}$. We compute class probabilities using a cosine classifier,

$$p_j(\mathbf{x}) = \frac{\exp(\beta \cos(\mathbf{v}, \mathbf{u}_j))}{\sum_{k \in \mathcal{C}^{\le t}} \exp(\beta \cos(\mathbf{v}, \mathbf{u}_k))}, \tag{7}$$

where $\mathbf{u}_j$ denotes the classifier weight for class j and $\beta > 0$ is a scaling factor. We first train OPD on base-session images using $\mathcal{L}_{\text{OPD}}$ and freeze it thereafter. PAM, the visual prompts, and the classifier weights for all seen classes are jointly optimized using cross-entropy in both base and incremental sessions. The visual prompts are shared across classes and sessions and carried over from the preceding session.

Table 1: Accuracy (%) on CUB-200 benchmark. **Avg.:**average accuracy over all sessions. **Bold:** the best performance.

Method	Venue	Base	Incremental sessions										Avg. ↑
		S0	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	
Fine-Tuning + Proto	–	84.21	66.43	25.00	25.44	16.19	4.58	1.42	1.49	3.62	5.50	3.79	21.60
iCaRL (Rebuffi et al., 2017)	CVPR'17	82.43	79.32	68.74	59.93	61.62	59.91	57.83	57.34	55.42	52.73	55.92	63.83
RDI (Zhou et al., 2024)	IJCAI'24	80.13	76.55	73.21	69.37	67.83	65.74	64.91	63.37	61.43	61.41	60.20	67.65
CEC (Zhang et al., 2021)	CVPR'21	81.82	79.53	78.42	75.54	76.31	74.83	74.41	74.62	74.23	73.91	73.84	76.13
FOSTER (Wang et al., 2022a)	ECCV'22	85.02	83.43	77.41	71.52	69.93	66.34	65.52	63.07	62.92	62.03	60.42	69.78
FACT (Zhou et al., 2022)	CVPR'22	84.32	81.23	79.14	75.13	75.42	73.31	72.43	72.52	71.41	71.12	70.91	75.18
CLOM (Zou et al., 2022)	NeurIPS'22	83.28	81.85	79.61	77.79	76.34	74.64	73.62	72.82	71.24	71.33	70.50	75.73
NC-FSCIL (Yang et al., 2023b)	ICLR'23	83.52	80.92	80.14	77.83	77.81	76.96	76.72	74.78	74.18	73.92	73.80	77.32
WaRP (Kim et al., 2023)	ICLR'23	82.74	80.21	79.06	77.80	77.78	76.81	76.82	74.61	74.13	74.02	73.36	77.03
TEEN (Wang et al., 2023)	NeurIPS'23	84.03	81.52	80.91	78.34	78.32	77.24	77.13	75.42	75.51	75.13	75.61	78.11
Comp-FSCIL (Zou et al., 2024)	ICML'24	83.67	81.73	79.03	78.04	77.73	75.52	74.32	74.55	73.35	73.15	72.80	76.72
Yourself (Tang et al., 2024)	ECCV'24	82.31	80.65	79.85	76.96	77.08	74.85	74.99	75.40	74.29	74.69	74.38	76.86
PriViLege (Park et al., 2024)	CVPR'24	82.21	81.25	80.45	77.76	77.78	75.95	75.69	76.00	75.19	75.19	75.08	77.50
ConCM (Wang et al., 2026)	ICLR'26	72.08	49.7	53.77	52.07	51.56	50.51	48.95	47.93	47.01	47.41	47.94	51.72
CompSLOT (Liao et al., 2026)	ICLR'26	84.36	82.56	81.98	81.06	80.78	78.90	78.74	78.66	77.89	77.99	78.05	80.09
L2P (Wang et al., 2022c)	CVPR'22	82.47	81.23	79.01	76.89	76.21	74.73	74.19	74.11	72.73	73.02	73.67	76.21
DualPrompt (Wang et al., 2022b)	ECCV'22	83.51	82.27	80.93	79.57	78.63	77.09	76.31	77.03	75.79	76.17	76.53	78.53
CODA-Prompt (Smith et al., 2023)	CVPR'23	79.61	78.12	76.42	75.68	75.02	73.19	72.58	72.81	72.07	72.49	72.97	74.63
LGSP-Prompt (Jiang et al., 2025)	ICCV'25	85.72	84.31	83.21	81.33	81.80	80.33	79.98	80.09	79.18	79.74	79.72	81.39
SEC-Prompt (Liu & Yang, 2025)	CVPR'25	87.50	86.61	85.51	84.62	84.94	84.02	83.71	84.24	83.81	84.03	84.82	84.87
PDAM	Ours	**90.96**	**89.84**	**90.01**	**88.90**	**88.02**	**87.53**	**87.39**	**87.14**	**86.07**	**86.18**	**85.47**	**87.96**
Better than second		+3.46	+3.23	+4.50	+4.28	+3.08	+3.51	+3.68	+2.90	+2.26	+2.15	+0.65	+3.09

4 EXPERIMENTS

4.1 SETUP

Benchmarks and compared methods. We evaluate PDAM on three fine-grained FSCIL benchmarks, CUB-200 (Wah et al., 2011), iNF200 (D'Alessandro et al., 2023), and FGVC-Aircraft (Maji et al., 2013), covering birds, fungi, and aircraft, respectively. These benchmarks require distinguishing visually similar categories from limited labeled examples, allowing us to assess the effectiveness of primitive-guided feature modulation. We follow the class splits and session configurations of LGSP-Prompt (Jiang et al., 2025). Our comparisons include representative prompt-based CIL methods, including DualPrompt (Wang et al., 2022b), CODA-Prompt (Smith et al., 2023), LGSP-Prompt (Jiang et al., 2025), and SEC-Prompt (Liu & Yang, 2025), alongside FSCIL methods such as TEEN (Wang et al., 2023), Comp-FSCIL (Zou et al., 2024), and PriViLege (Park et al., 2024).

Implementation details. We adopt a pretrained DINOv2 backbone on LVD-142M (Oquab et al., 2023) as the image encoder. Following VPT-Deep, we insert 5 learnable prompt tokens into each Transformer block from blocks 5–10 for CUB-200 and blocks 6–10 for iNF200 and FGVC-Aircraft. The pretrained backbone remains frozen throughout training. We set the number of slots to 8 and train OPD on base-class images for 200 epochs with a learning rate of 0.002. OPD is then frozen during subsequent training. We set $d_r = 32$ and preserve the feature dimension $D = 768$ in f_θ. The cosine classifier uses a logit scaling factor of $\beta = 16$. PAM, the visual prompts, and the classifier are jointly trained in the base session, then fine-tuned for 5 epochs per incremental session at a learning rate of 0.003. All experiments use one NVIDIA GeForce RTX 4090 GPU.

4.2 RESULTS

Quantitative results. In this section, we compare PDAM with state-of-the-art methods on three fine-grained FSCIL benchmarks. Tables 1, 2, and 3 report the results on CUB-200, FGVC-Aircraft, and iNF200, respectively. The proposed method outperforms previous SOTA methods across all datasets, with the average accuracy gains across all sessions of 3.09% on CUB200, 10.36% on FGVC-Aircraft and 6.79% on iNF200. Within individual incremental sessions, its largest margins over the best competing results reach 4.50%, 13.96%, and 10.41% on the three benchmarks. Additional results on miniImageNet and CIFAR-100 are provided in the Appendix.

Table 2: Accuracy (%) on the FGVC-Aircraft benchmark. **Avg.:** average accuracy over all sessions. **Bold:** the best performance.

Method	Venue	Base	Incremental sessions										Avg.↑
		S0	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	
Yourself (Tang et al., 2024)	ECCV'24	26.29	20.37	17.05	14.82	13.16	11.87	10.83	9.97	9.24	8.62	8.08	13.66
Approximation (Wang et al., 2024)	ECCV'24	21.67	19.48	18.90	17.57	16.50	16.68	15.82	15.21	14.20	13.85	12.97	16.62
PriViLege (Park et al., 2024)	CVPR'24	22.75	20.79	21.85	20.96	20.27	22.32	21.53	21.67	20.43	20.18	20.04	21.16
LGSP-Prompt (Jiang et al., 2025)	ICCV'25	27.67	24.82	26.30	27.38	25.25	26.76	26.37	25.49	25.50	24.83	23.87	25.82
SEC-Prompt (Liu & Yang, 2025)	CVPR'25	43.22	39.28	36.70	36.52	35.88	35.40	36.05	32.33	31.00	27.29	26.97	34.60
CompSLOT (Liao et al., 2026)	ICLR'26	37.25	30.97	27.35	27.79	26.92	28.56	27.49	27.96	21.77	22.20	22.42	27.33
PDAM	Ours	**49.84**	**48.87**	**50.65**	**50.48**	**47.99**	**43.68**	**42.95**	**40.45**	**40.17**	**41.16**	**38.37**	**44.96**
Better than second		+6.62	+9.59	+13.95	+13.96	+12.11	+8.28	+6.90	+8.12	+9.17	+13.87	+11.40	+10.36

Table 3: Accuracy (%) on the iNF200 benchmark. **Avg.:** average accuracy over all sessions. **Bold:** the best performance.

Method	Venue	Base	Incremental sessions										Avg.↑
		S0	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	
Yourself (Tang et al., 2024)	ECCV'24	61.90	55.73	50.50	45.31	40.93	35.07	30.75	26.18	22.72	19.47	16.10	36.97
Approximation (Wang et al., 2024)	ECCV'24	66.30	60.18	57.25	53.62	50.43	47.67	44.88	42.41	40.11	38.32	36.75	48.90
PriViLege (Park et al., 2024)	CVPR'24	63.10	60.73	59.50	58.31	57.93	56.07	53.75	52.18	51.72	49.47	48.10	55.53
LGSP-Prompt (Jiang et al., 2025)	ICCV'25	67.00	63.27	62.83	60.92	59.57	57.67	56.19	53.88	53.83	50.05	48.60	57.62
SEC-Prompt (Liu & Yang, 2025)	CVPR'25	76.20	69.91	67.92	67.69	67.86	67.47	67.38	65.82	63.83	63.37	61.35	67.16
CompSLOT (Liao et al., 2026)	ICLR'26	71.40	65.27	65.58	66.23	67.21	67.33	67.00	66.41	66.50	64.74	60.35	66.18
PDAM	Ours	**80.20**	**78.64**	**78.33**	**77.69**	**75.93**	**75.33**	**74.88**	**70.88**	**69.11**	**67.21**	**65.20**	**73.95**
Better than second		+4.00	+8.73	+10.41	+10.00	+8.07	+7.86	+7.50	+4.47	+2.61	+2.47	+3.85	+6.79

Qualitative results. (1) As shown in Fig. 3(a), OPD captures meaningful visual components in novel classes without further training, supporting the transferability of the decomposition capability learned from base classes. (2) Fig. 3(b) shows that initially fragmented primitive assignments become more spatially coherent over successive iterations, illustrating the effect of iterative refinement. Additional attention visualizations and comparisons with the baseline are provided in the Appendix.

4.3 Analysis

Ablation study. As shown in Tab. 4, PW and CM exhibit a clear complementary effect. On CUB-200, using either module alone brings only marginal changes, while combining them improves the overall accuracy from 85.75% to 87.96%, with consistent gains on both base session and novel sessions. Similar improvements are observed on FGVC-Aircraft, where the overall accuracy increases from 40.21% to 44.96%, particularly improving novel-session accuracy from 39.38% to 44.48%. These results demonstrate that PW and CM work synergistically to better exploit primitive knowledge for both base session and novel sessions.

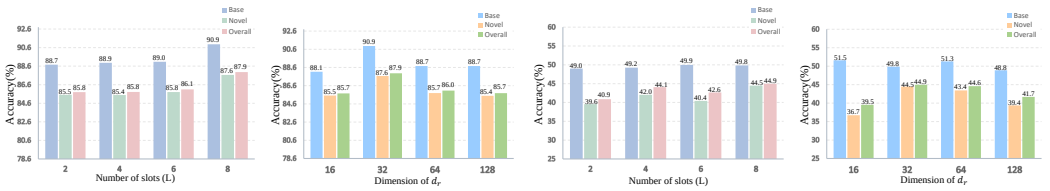
Table 4: Ablation of primitive weighting (PW) and channel modulation (CM). O, B, and N denote overall, base-session, and average accuracy over incremental sessions (%), respectively.

PW	CM	CUB-200			FGVC-Aircraft		
		O↑	B↑	N↑	O↑	B↑	N↑
		85.75	88.29	85.50	40.21	48.57	39.38
✓		85.57	87.87	85.34	42.68	**54.46**	41.51
	✓	85.82	87.81	85.62	42.20	48.40	41.58
✓	✓	**87.96**	**90.96**	**87.66**	**44.96**	49.84	**44.48**

Analysis of Primitive Weighting Strategies. We compare soft weighting with hard selection, which retains the top-k primitives ranked by primitive weights in Tab. 5. On CUB-200, soft weighting achieves the best performance across all three metrics, improving average accuracy to 87.96%, compared with 86.92% and 86.91% for top-6 and top-7 selection, respectively. On FGVC-Aircraft, soft weighting achieves the highest average accuracy of 44.96%, although top-7 selection yields higher base-session accuracy and top-6 selection gives slightly higher incremental-

Table 5: Comparison of hard selection and soft weighting.

Strategy	CUB-200			FGVC-Aircraft		
	O↑	B↑	N↑	O↑	B↑	N↑
Top-6 selection	86.92	87.52	86.86	44.87	48.46	44.51
Top-7 selection	86.91	87.41	86.86	44.76	**52.30**	44.01
Soft weighting	**87.96**	**90.96**	**87.66**	**44.96**	49.84	**44.48**



(a) CUB-200 (b) FGVC Aircraft

Figure 4: Effect of the number of L and d_r on (a) CUB-200 and (b) FGVC-Aircraft.

session accuracy. By continuously adjusting primitive contributions rather than discarding unselected primitives, soft weighting may better preserve complementary visual evidence.

Analysis of pretrained backbones. To assess whether the effectiveness of PDAM stems solely from the use of DINOv2, we replace its backbone with the ImageNet-21K-pretrained ViT-B/16 used by the compared methods, while keeping the remaining components unchanged. As shown in Tab. 6, PDAM still outperforms these methods with the same ViT backbone, indicating that its advantage is not solely attributable to DINOv2. We further evaluate LGSP-Prompt and SEC-Prompt with DINOv2, where PDAM also maintains its lead. Together, these comparisons support the effectiveness of primitive-guided adaptation across both pretrained backbones.

Table 6: Analysis of pretrained backbones on CUB-200.

	ViT-B/16 (IN-21K)			DINOv2 ViT-B/14		
Method	O↑	B↑	N↑	O↑	B↑	N↑
CODA-Prompt	74.63	79.61	74.13	77.27	86.51	76.35
LGSP-Prompt	81.39	85.72	80.96	83.98	86.24	83.74
SEC-Prompt	84.87	87.50	84.63	84.93	88.59	84.57
Ours	**85.79**	**88.86**	**85.48**	**87.96**	**90.96**	**87.66**

Joint versus decoupled training. Tab. 7 compares training OPD jointly with classification against learning it independently. Decoupled training improves overall accuracy by 2.15% on CUB-200, and 3.69% on FGVC-Aircraft, respectively, with higher average accuracy across incremental sessions on both datasets. These results support separating primitive discovery from classification, consistent with the motivation discussed in Remark 1.

Table 7: Comparison of primitive learning strategies.

Strategy	CUB-200			FGVC		
	O↑	B↑	N↑	O↑	B↑	N↑
Joint learning	85.81	88.82	85.51	41.27	**52.25**	40.05
Decoupled learning	**87.96**	**90.96**	**87.66**	**44.96**	49.84	**44.48**

Analysis of hyper-parameter. Fig. 4 evaluates the effects of L and d_r. Increasing L from 2 to 8 improves overall accuracy from 85.8% to 87.9% on CUB-200 and from 40.9% to 44.9% on FGVC-Aircraft, although the gains are not monotonic. This suggests that additional primitives can support a more effective visual decomposition. For the projection dimension, $d_r = 32$ achieves the best overall accuracy on both datasets, while larger dimensions provide no further benefit.

5 Conclusion

In this work, we propose PDAM, an FSCIL framework that transfers primitive discovery from base classes to guide feature adaptation for novel categories. OPD learns to decompose images into visual primitives through self-supervised feature reconstruction and remains frozen during subsequent classification training. Building on this input-specific decomposition, PAM combines primitive-level weighting with primitive-conditioned channel modulation to support fine-grained discrimination under limited supervision. Experiments on three fine-grained FSCIL benchmarks demonstrate consistent improvements over existing methods, supporting the effectiveness of PDAM.

REFERENCES

Adil Kaan Akan and Yücel Yemez. Slot-guided adaptation of pre-trained diffusion models for object-centric learning and compositional generation. In *International Conference on Learning Representations*, 2025.

Magdalena Biesialska, Katarzyna Biesialska, and Marta R Costa-Jussa. Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th international conference on computational linguistics*, pp. 6523–6541, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Sarthak Chandra, Sugandha Sharma, Rishidev Chaudhuri, and Ila Fiete. Episodic and associative memory from spatial scaffolds in the hippocampus. *Nature*, 638(8051):739–751, 2025.

Chaoqi Chen, Luyao Tang, and Hui Huang. Reconstruct and match: Out-of-distribution robustness via topological homogeneity. *Advances in Neural Information Processing Systems*, 37:125588–125607, 2024.

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

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

Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pp. 1597–1600. IEEE, 2017.

Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. PODNet: Pooled outputs distillation for small-tasks incremental learning. In *European Conference on Computer Vision*, pp. 86–102. Springer, 2020. doi: 10.1007/978-3-030-58565-5_6.

Matteo Farina, Massimiliano Mancini, Giovanni Iacca, and Elisa Ricci. Rethinking few-shot adaptation of vision-language models in two stages. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29989–29998, 2025.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1126–1135, 2017.

Nathaniel R Goodman. Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction). *The Annals of mathematical statistics*, 34(1):152–177, 1963.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in neural information processing systems*, 37:59532–59569, 2024.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

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

Yongwei Jiang, Yixiong Zou, Yuhua Li, and Ruixuan Li. Revisiting pool-based prompt learning for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1303–1313, 2025.

Ioannis Kakogeorgiou, Spyros Gidaris, Konstantinos Karantzalos, and Nikos Komodakis. SPOT: Self-training with patch-order permutation for object-centric learning with autoregressive transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22776–22786, 2024.

Do-Yeon Kim, Dong-Jun Han, Jun Seo, and Jaekyun Moon. Warping the space: Weight space rotation for class-incremental few-shot learning. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=kPLzOfPfA2l.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision*, pp. 614–629. Springer, 2016. doi: 10.1007/978-3-319-46493-0_37.

Weiduo Liao, Fei Han, Hisao Ishibuchi, Qingfu Zhang, and Ying Wei. Plug-and-play compositionality for boosting continual learning with foundation models. In *International Conference on Learning Representations*, volume 2026, pp. 82486–82519, 2026.

Chenxi Liu, Zhenyi Wang, Tianyi Xiong, Ruibo Chen, Yihan Wu, Junfeng Guo, and Heng Huang. Few-shot class incremental learning with attention-aware self-adaptive prompt. In *Computer Vision – ECCV 2024*, volume 15139 of *Lecture Notes in Computer Science*, pp. 1–18. Springer, 2024. doi: 10.1007/978-3-031-73004-7_1.

Ye Liu and Meng Yang. Sec-prompt: Semantic complementary prompting for few-shot class-incremental learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25643–25656. IEEE, 2025.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Keon-Hee Park, Kyungwoo Song, and Gyeong-Moon Park. Pre-trained vision and language transformers are few-shot incremental learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23881–23890, 2024. doi: 10.1109/CVPR52733.2024.02254.

Hang Qi, Matthew Brown, and David G. Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5822–5830, 2018.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5533–5542, 2017. doi: 10.1109/CVPR.2017.587.

Chinmay Savadikar, Michelle Dai, and Tianfu Wu. CHEEM: Continual learning by reuse, new, adapt and skip – a hierarchical exploration-exploitation approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25066–25076, 2026.

Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogerio Feris, Raja Giryes, and Alex M. Bronstein. Delta-encoder: An effective sample synthesis method for few-shot object recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2850–2860, 2018.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. *arXiv preprint arXiv:2209.14860*, 2022.

Guangyuan Shi, Jiaxin Chen, Wenlong Zhang, Li-Ming Zhan, and Xiao-Ming Wu. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. *Advances in neural information processing systems*, 34:6747–6761, 2021.

Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18181–18196, 2022.

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

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4077–4087, 2017.

Luyao Tang, Kunze Huang, Chaoqi Chen, Yuxuan Yuan, Chenxin Li, Xiaotong Tu, Xinghao Ding, and Yue Huang. Dissecting generalized category discovery: Multiplex consensus under self-deconstruction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 297–307. IEEE, 2025a.

Luyao Tang, Yuxuan Yuan, Chaoqi Chen, Zeyu Zhang, Yue Huang, and Kun Zhang. Ocrt: Boosting foundation models in the open world with object-concept-relation triad. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25422–25433. IEEE, 2025b.

Yu-Ming Tang, Yi-Xing Peng, Jingke Meng, and Wei-Shi Zheng. Rethinking few-shot class-incremental learning: Learning from yourself. In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pp. 108–128. Springer, 2024. doi: 10.1007/978-3-031-73030-6_7.

Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12180–12189. IEEE, 2020.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. FOSTER: Feature boosting and compression for class-incremental learning. In *Computer Vision – ECCV 2022*, volume 13685 of *Lecture Notes in Computer Science*, pp. 398–414. Springer, 2022a. doi: 10.1007/978-3-031-19806-9_23.

Qi-Wei Wang, Da-Wei Zhou, Yi-Kai Zhang, De-Chuan Zhan, and Han-Jia Ye. Few-shot class-incremental learning via training-free prototype calibration. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15060–15076, 2023. doi: 10.52202/075280-0661.

Qinzhe Wang, Zixuan Chen, Keke Huang, Xiu Su, Chunhua Yang, and Chang Xu. Consistency-driven calibration and matching for few-shot class incremental learning. In *International Conference on Learning Representations*, volume 2026, pp. 1255–1279, 2026.

Xuan Wang, Zhong Ji, Xiyao Liu, Yanwei Pang, and Jungong Han. On the approximation risk of few-shot class-incremental learning. In *Computer Vision – ECCV 2024*, volume 15109 of *Lecture Notes in Computer Science*, pp. 162–178. Springer, 2024. doi: 10.1007/978-3-031-72983-6_10.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7278–7286, 2018.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. DualPrompt: Complementary prompting for rehearsal-free continual learning. In *Computer Vision – ECCV 2022*, volume 13686 of *Lecture Notes in Computer Science*, pp. 631–648. Springer, 2022b. doi: 10.1007/978-3-031-19809-0_36.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 139–149, 2022c.

Shipeng Yan, Jiangwei Xie, and Xuming He. DER: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3014–3023, 2021.

Haoyuan Yang, Xiaoou Li, Jiaming Lv, Xianjun Cheng, Qilong Wang, and Peihua Li. ImagineFSL: Self-supervised pretraining matters on imagined base set for VLM-based few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30020–30031, 2025.

Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In *International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=y5W8tpojhtJ.

Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip H. S. Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023b. URL https://openreview.net/forum?id=y5W8tpojhtJ.

Yuxuan Yuan, Luyao Tang, Yixin Chen, Chaoqi Chen, Yue Huang, and Xinghao Ding. Asgs: Single-domain generalizable open-set object detection via adaptive subgraph searching. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20911–20921. IEEE, 2025.

Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12455–12464, 2021. doi: 10.1109/CVPR46437.2021.01227.

Xin Zhang, Liang Bai, Guanchao Wang, and Xian Yang. Exemplar-free class incremental learning via preserving class-discriminative structure. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17979–17988, 2026.

Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, and De-Chuan Zhan. Forward compatible few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9046–9056, 2022.

Haichen Zhou, Yixiong Zou, Ruixuan Li, Yuhua Li, and Kui Xiao. Delve into base-novel confusion: Redundancy exploration for few-shot class-incremental learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5635–5643, 2024. doi: 10.24963/ijcai.2024/623.

Yixiong Zou, Shanghang Zhang, Yuhua Li, and Ruixuan Li. Margin-based few-shot class-incremental learning with class-level overfitting mitigation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27267–27279, 2022. doi: 10.52202/068431-1977.

Yixiong Zou, Shanghang Zhang, Haichen Zhou, Yuhua Li, and Ruixuan Li. Compositional few-shot class-incremental learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62964–62977. PMLR, 2024. URL https://proceedings.mlr.press/v235/zou24c.html.

Appendix

A AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing of the final manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

B ADDITIONAL RELATED WORKS

B.1 FEW-SHOT LEARNING

Few-shot learning (FSL) aims to recognize novel classes from limited labeled examples. Existing methods mainly approach this problem from the perspectives of metric learning (Snell et al., 2017; Qi et al., 2018), meta-learning (Finn et al., 2017; Nichol et al., 2018), and data augmentation (Wang et al., 2018; Schwartz et al., 2018). Recently, pretrained vision-language and generative models have further advanced FSL by providing transferable semantic and visual priors. For example, ImagineFSL (Yang et al., 2025) exploits synthetic images as an imagined base set for representation transfer, while 2SFS (Farina et al., 2025) performs parameter-efficient adaptation of vision-language models under few-shot supervision. Nevertheless, conventional FSL generally focuses on adapting to a fixed set of novel classes, without considering the sequential arrival of new classes and the preservation of previously acquired knowledge, which are central challenges in FSCIL.

B.2 CLASS INCREMENTAL LEARNING

Class-incremental learning (CIL) aims to sequentially acquire new classes while retaining recognition performance on previously learned ones. Existing approaches address catastrophic forgetting through several complementary strategies. Replay-based methods (Rebuffi et al., 2017) retain representative old-class examples for subsequent training. Regularization-based methods (Kirkpatrick et al., 2017) constrain updates to important parameters, while knowledge distillation (Li & Hoiem, 2016; Douillard et al., 2020) preserves previous predictions or feature representations. Architecture-based approaches (Yan et al., 2021) expand model capacity to accommodate new classes while retaining previously learned features. With the adoption of pretrained models, recent research increasingly explores how to adapt transferable representations without disrupting existing knowledge. CHEEM (Savadikar et al., 2026) dynamically reuses, adds, adapts, or skips network components, while Zhang et al. (Zhang et al., 2026) preserve intra-class distributions and inter-class geometric relationships to mitigate forgetting without exemplar replay. Beyond parameter and representation preservation, CompSLOT (Liao et al., 2026) introduces compositional concept learning and distills similarities between primitive representations into classifier outputs. Despite these advances, extending CIL to few-shot settings requires reliable novel-class adaptation from scarce examples while preserving earlier knowledge.

B.3 VISUAL PROMPT TUNING

Visual prompt tuning (VPT) (Jia et al., 2022) adapts pretrained vision transformers by introducing learnable prompt tokens while keeping the backbone frozen. VPT-Shallow inserts prompts at the input layer, whereas VPT-Deep introduces prompts at multiple layers. In continual learning, L2P (Wang et al., 2022c) selects input-dependent prompts from a learnable pool, DualPrompt (Wang et al., 2022b) combines task-invariant and task-specific prompts, and CODA-Prompt (Smith et al., 2023) constructs prompts through attention-weighted combinations of learnable components. For FSCIL, ASP (Liu et al., 2024) learns shared and adaptive task-specific prompts, SEC-Prompt (Liu & Yang, 2025) captures complementary discriminative and non-discriminative information, and LGSP-Prompt (Jiang et al., 2025) constructs spatial prompts from local features and global frequency-domain representations. These approaches improve adaptation through prompt design and selection. Our method complements this direction by using visual primitives to modulate patch features at regional and channel levels, emphasizing discriminative evidence for subsequent VPT-Deep adaptation.

C THEORETICAL ANALYSIS

We analyze three structural properties of primitive-guided adaptive modulation: the spatial interpretation of its channel relation descriptor, invariance to primitive ordering, and bounded feature modulation. These properties follow from the formulation of PAM. The transferability of OPD to novel categories is evaluated empirically; the analysis below does not assume that reconstruction alone guarantees semantic decomposition or compositional generalization.

Notation. Let $\mathbf{z} \in \mathbb{R}^{N_p \times D}$ denote the patch features, and let $\mathbf{M} \in \mathbb{R}^{L \times N_p}$ contain the primitive masks. The masks and primitive weights satisfy,

$$m_{ln} \geq 0, \qquad \sum_{l=1}^{L} m_{ln} = 1, \qquad w_l \geq 0, \qquad \sum_{l=1}^{L} w_l = 1. \tag{8}$$

The channel adjustments form $\mathbf{Q} = [\mathbf{g}_1, \ldots, \mathbf{g}_L]^\top \in (-1, 1)^{L \times D}$. PAM computes:

$$\mathbf{G} = \mathbf{M}^\top \operatorname{Diag}(\mathbf{w})\mathbf{Q}, \qquad \tilde{\mathbf{z}} = \mathbf{z} + \lambda(\mathbf{G} \odot \mathbf{z}). \tag{9}$$

Proposition 1 (Spatial interpretation of the relation descriptor). We define the centering matrix:

$$\mathbf{C} = \mathbf{I}_{N_p} - \frac{1}{N_p}\mathbf{1}\mathbf{1}^\top. \tag{10}$$

For $\epsilon > 0$, the normalization can be written as:

$$\mathcal{N}_\epsilon(\mathbf{v}) = \frac{\mathbf{C}\mathbf{v}}{\|\mathbf{C}\mathbf{v}\|_2 + \epsilon}. \tag{11}$$

For nonconstant $\mathbf{v}$, we define:

$$\eta_{\mathbf{v}} = \frac{\|\mathbf{C}\mathbf{v}\|_2}{\|\mathbf{C}\mathbf{v}\|_2 + \epsilon}. \tag{12}$$

If the channel response $\mathbf{a}_d$ and primitive mask $\mathbf{m}_l$ are both nonconstant, their relation descriptor satisfies:

$$r_{ld} = \eta^2_{\mathbf{a}_d} + \eta^2_{\mathbf{m}_l} - 2\eta_{\mathbf{a}_d}\eta_{\mathbf{m}_l}\rho(\mathbf{a}_d, \mathbf{m}_l), \tag{13}$$

where

$$\rho(\mathbf{a}_d, \mathbf{m}_l) = \frac{(\mathbf{C}\mathbf{a}_d)^\top(\mathbf{C}\mathbf{m}_l)}{\|\mathbf{C}\mathbf{a}_d\|_2\|\mathbf{C}\mathbf{m}_l\|_2} \tag{14}$$

is their spatial Pearson correlation. Moreover, $0 \leq r_{ld} \leq 4$.

Proof. The normalized vector can be expressed as:

$$\mathcal{N}_\epsilon(\mathbf{v}) = \eta_{\mathbf{v}} \frac{\mathbf{C}\mathbf{v}}{\|\mathbf{C}\mathbf{v}\|_2}. \tag{15}$$

Substituting this expression into:

$$r_{ld} = \|\mathcal{N}_\epsilon(\mathbf{a}_d) - \mathcal{N}_\epsilon(\mathbf{m}_l)\|_2^2 \tag{16}$$

and expanding the squared norm yields Eq. 13. Since $\|\mathcal{N}_\epsilon(\mathbf{v})\|_2 \leq 1$, the triangle inequality gives

$$r_{ld} \leq \left(\|\mathcal{N}_\epsilon(\mathbf{a}_d)\|_2 + \|\mathcal{N}_\epsilon(\mathbf{m}_l)\|_2\right)^2 \leq 4. \tag{17}$$

Nonnegativity follows from the squared norm.

For fixed nonconstant responses, as $\epsilon \to 0$,

$$r_{ld} \longrightarrow 2 - 2\rho(\mathbf{a}_d, \mathbf{m}_l). \tag{18}$$

Thus, when both centered norms are large relative to ϵ, the descriptor approximately measures spatial disagreement through correlation. Centering provides exact invariance to uniform offsets:

$$\mathcal{N}_\epsilon(\mathbf{v} + b\mathbf{1}) = \mathcal{N}_\epsilon(\mathbf{v}), \qquad b \in \mathbb{R}. \tag{19}$$

Positive rescaling is not exactly invariant for finite ϵ, because:

$$\mathcal{N}_\epsilon(a\mathbf{v}) = \frac{a\mathbf{C}\mathbf{v}}{a\|\mathbf{C}\mathbf{v}\|_2 + \epsilon}, \qquad a > 0. \tag{20}$$

Its influence becomes small when the centered norms before and after rescaling are sufficiently large relative to ϵ. For constant responses, the normalization remains well-defined and returns zero, although the correlation interpretation does not apply. Spatial correspondence does not itself establish class relevance; the shared mapping f_θ learns how these relations should affect channel modulation.

Proposition 2 (Invariance to primitive ordering). Assume that primitive weights and channel adjustments are computed using mappings shared across primitives, without index-specific parameters. For any permutation matrix $\mathbf{P} \in \{0,1\}^{L\times L}$, reordering the primitives leaves $\mathbf{G}$ and $\tilde{\mathbf{z}}$ unchanged.

Proof. Reordering the primitives gives

$$\mathbf{M}' = \mathbf{P}\mathbf{M}, \qquad \mathbf{w}' = \mathbf{P}\mathbf{w}, \qquad \mathbf{Q}' = \mathbf{P}\mathbf{Q}. \tag{21}$$

The latter two identities follow from the shared content-based mappings and the permutation equivariance of softmax. Using

$$\mathrm{Diag}(\mathbf{P}\mathbf{w}) = \mathbf{P}\,\mathrm{Diag}(\mathbf{w})\mathbf{P}^\top \tag{22}$$

and $\mathbf{P}^\top\mathbf{P} = \mathbf{I}_L$, we obtain

$$\begin{aligned}\mathbf{G}' &= (\mathbf{P}\mathbf{M})^\top \mathrm{Diag}(\mathbf{P}\mathbf{w})(\mathbf{P}\mathbf{Q}) \\ &= \mathbf{M}^\top\mathbf{P}^\top\mathbf{P}\,\mathrm{Diag}(\mathbf{w})\mathbf{P}^\top\mathbf{P}\mathbf{Q} \\ &= \mathbf{M}^\top\,\mathrm{Diag}(\mathbf{w})\mathbf{Q} = \mathbf{G}.\end{aligned} \tag{23}$$

Since reordering does not change the patch features $\mathbf{z}$, the modulated features $\tilde{\mathbf{z}} = \mathbf{z} + \lambda(\mathbf{G}\odot\mathbf{z})$ also remain unchanged. Thus, PAM does not require a fixed semantic identity for each slot index. This property concerns reordering the same set of primitives; it does not imply invariance to changes in the decomposition itself.

Proposition 3 (Structured modulation and bounded feature changes). Let $\mathbf{M} \in \mathbb{R}^{L\times N_p}$, $\mathbf{w} \in \mathbb{R}^L$, and $\mathbf{Q} \in \mathbb{R}^{L\times D}$. Assume that $m_{ln} \geq 0$, $\sum_{l=1}^{L} m_{ln} = 1$, $w_l \geq 0$, $\sum_{l=1}^{L} w_l = 1$, and $|g_{ld}| \leq 1$, where g_{ld} denotes an entry of $\mathbf{Q}$. The modulation matrix $\mathbf{G} = \mathbf{M}^\top\,\mathrm{Diag}(\mathbf{w})\mathbf{Q}$ satisfies

$$\mathrm{rank}(\mathbf{G}) \leq \min\{L, N_p, D\}, \qquad \|\mathbf{G}\|_{\max} \leq \|\mathbf{w}\|_\infty \leq 1, \tag{24}$$

where $\|\mathbf{G}\|_{\max} = \max_{n,d}|G_{nd}|$.

For any finite value of the learnable scalar $\lambda \in \mathbb{R}$, the modulated features $\widetilde{\mathbf{z}} = \mathbf{z} + \lambda(\mathbf{G}\odot\mathbf{z})$ satisfy

$$\|\widetilde{\mathbf{z}} - \mathbf{z}\|_F \leq |\lambda|\,\|\mathbf{w}\|_\infty\|\mathbf{z}\|_F. \tag{25}$$

Moreover,

$$\begin{aligned}\max\big\{0, 1 - |\lambda|\|\mathbf{w}\|_\infty\big\}\|\mathbf{z}\|_F \leq \|\widetilde{\mathbf{z}}\|_F, \\ \|\widetilde{\mathbf{z}}\|_F \leq \big(1 + |\lambda|\|\mathbf{w}\|_\infty\big)\|\mathbf{z}\|_F.\end{aligned} \tag{26}$$

Proof. The matrix-product rank inequality gives

$$\begin{aligned}\mathrm{rank}(\mathbf{G}) &\leq \min\big\{\mathrm{rank}(\mathbf{M}^\top), \mathrm{rank}(\mathbf{Q})\big\} \\ &\leq \min\{L, N_p, D\}.\end{aligned} \tag{27}$$

For each entry,

$$\begin{aligned}|G_{nd}| &= \left|\sum_{l=1}^{L} m_{ln}w_l g_{ld}\right| \\ &\leq \sum_{l=1}^{L} m_{ln}w_l|g_{ld}| \\ &\leq \|\mathbf{w}\|_\infty \sum_{l=1}^{L} m_{ln} = \|\mathbf{w}\|_\infty \leq 1.\end{aligned} \tag{28}$$

Consequently,

$$\begin{aligned}\|\widetilde{\mathbf{z}} - \mathbf{z}\|_F^2 &= \lambda^2 \sum_{n,d} G_{nd}^2 z_{nd}^2 \\ &\leq \lambda^2\|\mathbf{w}\|_\infty^2\|\mathbf{z}\|_F^2.\end{aligned} \tag{29}$$

Taking square roots proves: Eq. 25. The triangle and reverse triangle inequalities yield

$$\|\widetilde{\mathbf{z}}\|_F \le \|\mathbf{z}\|_F + \|\widetilde{\mathbf{z}} - \mathbf{z}\|_F, \tag{30}$$

$$\|\widetilde{\mathbf{z}}\|_F \ge \|\mathbf{z}\|_F - \|\widetilde{\mathbf{z}} - \mathbf{z}\|_F. \tag{31}$$

Combining these inequalities with Eq. 25 and the nonnegativity of the norm proves Eq. 26.

The rank bound concerns the modulation matrix $\mathbf{G}$, not necessarily the feature perturbation $\mathbf{G} \odot \mathbf{z}$. The feature-change bound depends explicitly on the learned magnitude $|\lambda|$. A positive lower multiplicative bound is obtained when $|\lambda|\|\mathbf{w}\|_\infty < 1$. Without an explicit constraint on λ, these results do not provide a uniform bound on feature changes throughout training or guarantee the absence of forgetting.

D MORE DETAILS

D.1 DATASET DETAILS

We evaluate PDAM on three fine-grained FSCIL benchmarks: CUB-200 (Wah et al., 2011), FGVC-Aircraft (Maji et al., 2013), and iNF200 (D'Alessandro et al., 2023), covering birds, aircraft, and fungi, respectively. Following prior work (Jiang et al., 2025), we focus on fine-grained recognition across these three domains, where distinguishing closely related categories requires capturing subtle visual differences. These benchmarks allow us to evaluate whether primitive-guided modulation helps identify discriminative visual cues under limited supervision. For CUB-200 and iNF200, the base session contains 100 classes with 30 and 50 labeled examples per class, respectively, followed by ten 10-way 5-shot incremental sessions. For FGVC-Aircraft, the base session contains 50 classes with 30 labeled examples per class, followed by ten 5-way 5-shot incremental sessions. We additionally evaluate PDAM on miniImageNet and CIFAR-100 to assess its applicability beyond fine-grained recognition. For both datasets, the base session contains 60 classes, followed by eight 5-way 5-shot incremental sessions.

D.2 MODEL ARCHITECTURE

Image encoder and visual prompts. We instantiate the image encoder f with a pretrained DINOv2 ViT-B/14 backbone. The extracted patch representations are $\mathbf{z} \in \mathbb{R}^{N_p \times D}$, with $N_p = 256$ and $D = 768$. Following VPT-Deep, we introduce 5 learnable prompt tokens per prompted Transformer block, each with dimensionality D. The visual prompts $\mathcal{P}$ are inserted into blocks 5–10 for CUB-200 and blocks 6–10 for FGVC-Aircraft and iNF200. All pretrained backbone parameters remain frozen throughout training.

Object-centric primitive discovery. OPD uses $L = 8$ slots with embedding dimension $D_s = 768$ and performs $R = 4$ Slot Attention iterations. The projection networks $\mathcal{K}_\beta$, $\mathcal{Q}_\gamma$, and $\mathcal{V}_\phi$ map the patch and slot representations into the shared D_s-dimensional attention space. Slot representations are iteratively refined through attention-based aggregation and GRU updates. The spatial broadcast decoder Φ_d broadcasts each refined slot to all N_p patch positions and incorporates positional embeddings. At each position, it predicts D feature values and one mask logit, producing an output of size $N_p \times (D + 1)$ per slot. A softmax across slots converts the mask logits into spatial masks $\mathbf{m}_l$, which combine the primitive-specific predictions into the reconstructed features $\hat{\mathbf{z}}$.

Primitive-guided adaptive modulation. For primitive-level weighting, the projections $\mathbf{W}_g$ and $\mathbf{W}_u$ map the global CLS representation and local primitive descriptors from $D = 768$ dimensions to a shared space of dimension $d_r = 32$. For channel-level modulation, each relation descriptor $\mathbf{r}_l \in \mathbb{R}^D$ is processed by the shared two-layer MLP f_θ, which preserves the feature dimension D. Its output is passed through a hyperbolic tangent to obtain the signed channel adjustments $\mathbf{g}_l$. These adjustments are combined with the primitive weights and spatial masks to construct $\mathbf{G} \in \mathbb{R}^{N_p \times D}$. The residual modulation coefficient λ is learnable. The resulting image representation $\mathbf{v}$ is classified using a cosine classifier with logit scaling factor $\beta = 16$.

Table 8: Accuracy (%) on the CIFAR-100 benchmark. **Avg.**: average accuracy over all sessions. **Bold**: the best performance among the reported results.

Method	Venue	Base	Incremental sessions								Avg.↑
		S0	S1	S2	S3	S4	S5	S6	S7	S8	
Fine-Tuning + Proto	–	91.3	73.9	41.6	40.4	41.6	13.9	16.4	8.7	5.1	37.0
CEC	CVPR'21	74.2	71.4	70.1	67.3	65.9	65.1	64.7	63.4	61.4	67.1
NC-FSCIL	ICLR'23	89.5	82.6	76.7	71.6	67.1	63.1	59.6	56.5	53.7	68.9
WaRP	ICLR'23	86.2	82.5	79.3	75.5	73.4	71.0	69.5	67.7	65.4	74.5
PriViLege	CVPR'24	90.8	89.3	88.9	87.5	87.8	87.3	87.5	87.1	86.0	88.0
L2P	CVPR'22	83.2	76.8	71.2	66.5	62.3	58.6	55.4	52.4	49.8	64.0
DualPrompt	ECCV'22	85.1	78.4	72.8	67.9	63.6	59.9	56.6	53.6	50.9	65.4
SEC-prompt	CVPR'25	92.0	90.8	90.6	89.1	**89.2**	89.1	89.2	**88.7**	**87.5**	89.5
PDAM	Ours	**93.1**	**91.5**	**90.7**	**89.2**	89.1	**89.4**	89.0	**88.7**	86.7	**89.7**
Δ *vs. best baseline*		+1.1	+0.7	+0.1	+0.1	-0.1	+0.3	-0.2	+0.0	-0.8	+0.2

Table 9: Accuracy (%) on the miniImageNet benchmark. **Avg.**: average accuracy over all sessions. **Bold**: the best performance among the reported results.

Method	Venue	Base	Incremental sessions								Avg.↑
		S0	S1	S2	S3	S4	S5	S6	S7	S8	
Fine-Tuning + Proto	–	93.7	87.1	73.5	50.3	26.4	7.3	23.5	29.7	9.9	44.6
CEC	CVPR'21	87.4	86.0	84.0	83.2	83.1	81.6	80.7	80.7	80.7	83.1
NC-FSCIL	ICLR'23	77.2	71.3	66.2	61.8	57.9	54.5	51.5	48.8	46.4	59.5
WaRP	ICLR'23	83.3	80.5	77.2	75.0	73.6	71.5	69.2	68.8	68.0	74.1
PriViLege	CVPR'24	96.7	96.5	95.7	95.5	95.5	94.9	94.3	94.2	94.1	95.3
L2P	CVPR'22	94.6	87.5	81.2	75.8	71.0	66.9	63.1	59.8	56.8	73.0
DualPrompt	ECCV'22	95.0	87.8	81.5	76.1	71.4	67.2	63.5	60.1	57.1	73.3
SEC-prompt	CVPR'25	97.2	**97.0**	95.6	**95.8**	**96.0**	**95.5**	**94.9**	**94.9**	**94.8**	**95.7**
PDAM	Ours	**97.3**	96.7	**96.0**	**95.8**	95.8	94.9	94.7	94.3	93.8	95.5
Δ *vs. best baseline*		+0.1	-0.3	+0.3	+0.0	-0.2	-0.6	-0.2	-0.6	-1.0	-0.2

Table 10: Comparison of FC and cosine classifiers on CUB-200. Accuracy (%) is reported for each session. Avg. denotes the average accuracy across all sessions.

Classifier	Base	Incremental sessions										Avg.↑
	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	
FC	84.79	84.18	83.81	83.09	82.16	80.34	80.65	81.27	80.97	81.49	81.76	82.23
Cosine	**90.96**	**89.84**	**90.01**	**88.90**	**88.02**	**87.53**	**87.39**	**87.14**	**86.07**	**86.18**	**85.47**	**87.96**

Table 11: Effect of fixed and learnable modulation strength λ on CUB-200. Accuracy (%) is reported for each session. Avg. denotes the average accuracy across all sessions.

Modulation strength	Base	Incremental sessions										Avg.↑
	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	
$\lambda = 0.4$	**91.23**	**90.26**	**90.75**	**89.02**	87.45	86.60	85.37	84.00	82.19	80.63	75.29	85.71
$\lambda = 0.6$	81.23	78.18	67.30	47.26	21.33	30.34	51.20	56.45	56.59	50.96	49.03	53.62
$\lambda = 0.8$	80.19	77.51	73.80	64.21	42.01	36.71	43.70	34.45	37.30	36.20	39.20	51.39
Learnable λ	90.96	89.84	90.01	88.90	**88.02**	**87.53**	**87.39**	**87.14**	**86.07**	**86.18**	**85.47**	**87.96**

E ADDITIONAL RESULTS

Results on miniImageNet and CIFAR-100. Following common FSCIL evaluation practice, we additionally conduct experiments on miniImageNet and CIFAR-100 to assess PDAM beyond fine-grained recognition. Tables 8 and 9 report accuracy across the base and eight incremental ses-

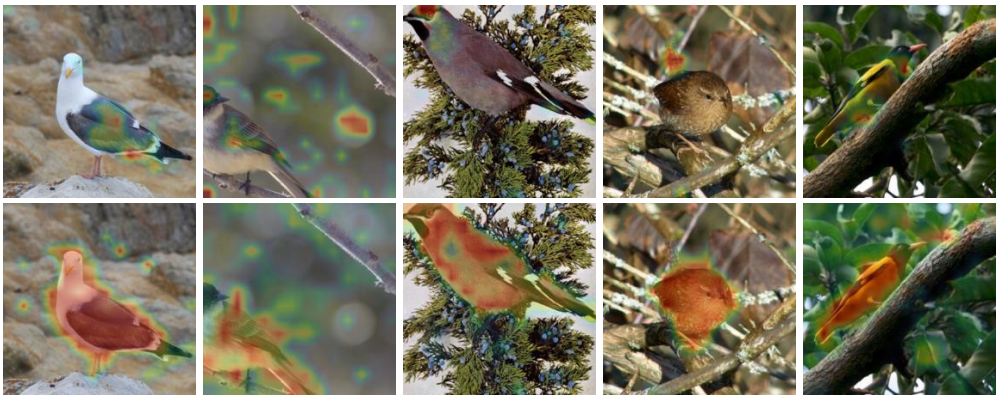
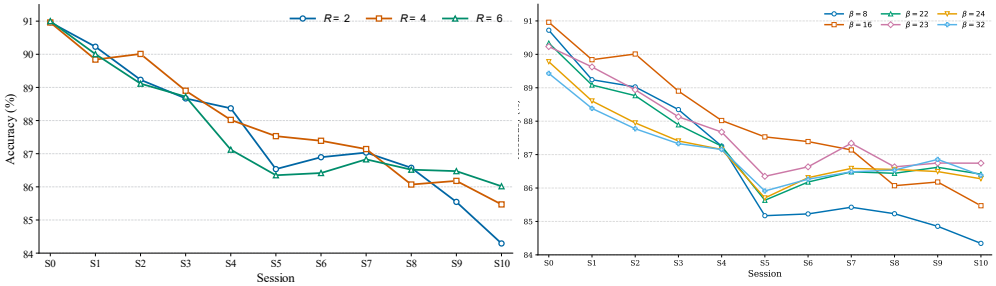


Figure 5: Attention map comparison between Comp-FSCIL (**top**) and PDAM (**bottom**). Each column shows the same image, with warmer colors indicating stronger responses.



(a) Number of Slot Attention iterations

(b) Number of logit scaling factor

Figure 6: Hyperparameter sensitivity of PDAM: session-wise accuracy with different (a) numbers of Slot Attention iterations R and (b) logit scaling factors β.

sions. On CIFAR-100, PDAM achieves an average accuracy of 89.7% and a final-session accuracy of 86.7%, while SEC-Prompt obtains 89.5% and 87.5%, respectively. The competitive performance across incremental sessions further demonstrates the applicability of primitive discovery and adaptive modulation to broader object categories. On miniImageNet, PDAM achieves an average accuracy of 95.5%, ranking second among the compared methods and closely following SEC-Prompt (95.7%). Although PDAM slightly trails SEC-Prompt on miniImageNet, it outperforms SEC-Prompt on all three fine-grained benchmarks. A possible explanation is that the broader category differences in miniImageNet reduce the benefit of modeling subtle local cues, while SEC-Prompt's semantic complementary prompting may better support generalization in this setting. In contrast, closely related fine-grained categories require more detailed visual distinctions, making primitive-guided modulation particularly beneficial.

Effect of the Cosine Classifier. To examine the influence of classifier design, we compare FC and cosine classifiers on CUB-200 while keeping the remaining components unchanged. As shown in Tab. 10, the cosine classifier consistently outperforms the FC classifier across all sessions, improving average accuracy from 82.23% to 87.96% and final-session accuracy from 81.76% to 85.47%. Unlike the FC classifier, the cosine classifier normalizes both image representations and classifier weights, making classification scores depend on their angular alignment rather than their magnitudes. The consistent improvements across base and incremental sessions support this classifier choice for the representations learned by PDAM.

Effect of Learning the Modulation Strength. We compare learnable modulation strength with fixed values of λ on CUB-200 in Tab. 11. Among the tested fixed settings, $\lambda = 0.4$ achieves

the highest average accuracy of 85.71%, whereas $\lambda = 0.6$ and $\lambda = 0.8$ yield substantially lower performance, indicating sensitivity to the fixed modulation strength. Although $\lambda = 0.4$ performs slightly better in the first four sessions, learnable λ achieves higher accuracy from S4 onward. These results support learning the modulation strength jointly with the other model parameters for more effective adaptation across incremental sessions.

Attention map visualizations. Fig. 5 compares the spatial attention patterns of Comp-FSCIL and PDAM across five bird images. Comp-FSCIL often activates only small portions of the bird or scattered background regions, leaving much of the foreground weakly highlighted. In contrast, PDAM produces more continuous responses over the birds' visible structures. In the first example, attention extends from a limited region near the wing and tail to the head, neck, and torso. The third and fourth examples similarly show broader coverage of the bird despite surrounding branches and foliage, while the fifth exhibits a more coherent response along the elongated body. In the second example, where the bird is partially outside the image, PDAM places stronger emphasis on its visible portion. Although some background responses remain, these visualizations suggest that primitive-guided modulation promotes more coherent attention to foreground structures.

Hyperparameter sensitivity analysis. Fig. 6 examines the effects of the number of Slot Attention iterations R and the logit scaling factor β. In Fig. 6(a), $R = 4$ achieves the highest accuracy in five sessions and the highest average accuracy of 87.96%, compared with 87.67% for $R = 2$ and 87.69% for $R = 6$. Although $R = 6$ performs best in the final session, increasing the number of iterations from four to six does not improve average performance, suggesting that additional refinement does not consistently benefit recognition. In Fig. 6(b), $\beta = 16$ performs best in the early sessions and achieves the highest average accuracy across all sessions. Larger scaling factors yield stronger results in several later sessions, but these gains do not offset their lower early-session accuracy. Meanwhile, $\beta = 8$ exhibits a pronounced performance drop after S4 and remains the weakest setting thereafter. These results support the choice of $R = 4$ and $\beta = 16$ based on average performance across sessions.