SUPERCLASS-GUIDED HIERARCHICAL PROMPTING FOR FEW-SHOT CLASS-INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-Shot Class-Incremental Learning (FSCIL) aims at continuously learning new-class knowledge from extremely limited samples while preserving the recognition capability on old classes, which suffers from the dual challenges of catastrophic forgetting and overfitting. While existing prompt-based FSCIL methods achieve promising results, they overlook the semantic relationships among classes and fail to explicitly exploit the shared visual-semantic patterns across related classes. In this paper, we propose a Superclass-guided Hierarchical Prompt learning framework (SH-Prompt) to address this issue. Specifically, our method organizes base classes into semantic superclasses via a large language model (LLM) and progressively assigns newly introduced classes to the established superclasses. Based on the resulting semantic structure, we design a key-based superclass prompt pool to adaptively incorporate superclass-shared prompts for each sample through a semantic-driven retrieval mechanism. Furthermore, we derive a superclass-guided instance-level prompt learning loss, which introduces structured regularization across superclass semantics to improve the inter-class separability. Extensive experimental results on three FSCIL benchmark datasets demonstrate that our proposed method outperforms the state-of-the-art. Code will be released publicly.

1 INTRODUCTION

In real-world applications, data distributions typically evolve continuously over time, requiring models to maintain stable recognition performance on previously learned classes while incrementally adapting to new classes. This fundamental need has driven the research on Class-Incremental Learning (CIL) (Rebuffi et al., 2017). Since the complete historical data is often unavailable for storage or repeated access, models tend to overwrite existing knowledge when learning new information, leading to the catastrophic forgetting problem (French, 1999). In more challenging practical scenarios, new classes come with only a handful of labeled samples due to prohibitive annotation costs or limited data collection. This gives rise to the Few-Shot Class-Incremental Learning (FSCIL) problem (Tao et al., 2020), where a model need to continuously learn novel-class knowledge from extremely limited annotated data while preserving its ability to recognize old classes. To address these issues, existing studies can be broadly categorized into two approaches. The first one (Chi et al., 2022; Shi et al., 2021; Zhu et al., 2021) focuses on leveraging the base model trained on data-abundant base classes to enable generalization to subsequent few-shot incremental tasks. The second one (Cheraghian et al., 2021; Dong et al., 2021; Kukleva et al., 2021; Wang et al., 2022a) concentrates on optimizing the training paradigm within the data-constrained incremental learning phase.

Prompt-based learning methods (Wang et al., 2022c;b; Smith et al., 2023) have emerged as a promising direction for FSCIL in recent years. These methods are typically built upon pre-trained Vision Transformer (ViT) (Dosovitskiy et al., 2021), where the backbone network is frozen and only a small number of learnable prompt parameters are introduced to encode task-specific knowledge. Existing methods construct a prompt pool and employ a key-query mechanism for prompt retrieval, storing knowledge from different tasks or classes in the prompt parameters. This design enables knowledge expansion without large-scale parameter updates, thus alleviating catastrophic forgetting. Nevertheless, most existing prompt-based methods primarily model classes independently, without explicitly exploiting the semantic relationships shared across classes. This overlooks the pervasive hierarchical structures among classes, where semantically related categories can share common visual and semantic patterns (e.g., “Golden Retriever” and “Husky” both belong to the “Dog” superclass).

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.

Modeling such shared patterns at the superclass level enables related classes to share prompts and learn from their common visual-semantic characteristics, benefiting both the recognition of base classes and the adaptation to newly introduced classes. A recent work (Vu et al., 2023) on few-shot instance segmentation explores semantic relationships among classes by constructing class hierarchies based on the similarity between base and novel classes. However, it mainly incorporates the hierarchical structure into the classification framework, rather than using it to learn shared prompts for related classes.

To this end, we approach the problem from the perspective of class relationship modeling and propose a Superclass-guided Hierarchical Prompt learning framework (SH-Prompt) which leverages the exploitable hierarchical semantic relations among classes as a structured prior. Specifically, we first organize the base classes into semantic superclasses via a large language model (LLM) and progressively assign newly introduced classes to the established superclasses during incremental learning. Based on the resulting superclass structure, we build a key-based superclass prompt pool, where each superclass corresponds to a set of shared prompt representations. A semantics-driven retrieval mechanism is then designed to achieve adaptive prompt retrieval for input samples, thereby introducing cross-class shared semantic information during novel class learning. In addition, to optimize representation quality across different semantic granularities, we design a superclass-guided instance-level prompt learning loss, which introduces structured constraints at the superclass semantic level. This design effectively improves inter-class separability and enhances the discriminability of the model. Through these designs, our model can achieve more effective knowledge transfer under limited data conditions, mitigating both overfitting and catastrophic forgetting, and thus achieving a better balance between stability and plasticity. The main contributions of this paper are summarized as follows:

- We propose SH-Prompt based on the superclass semantic hierarchy. A superclass prompt pool is built up to incorporate richer semantic knowledge during the incremental learning phase, enabling cross-category semantic sharing and stable knowledge transfer.
- We design a superclass-guided instance-level prompt learning loss, which optimizes representations at the superclass semantic granularity, thus simultaneously boosting the generalization capability and discriminability of the model.
- Extensive experiments on multiple FSCIL benchmark datasets demonstrate that the proposed method achieves a superior balance between novel class learning and old class retention, outperforming state-of-the-art approaches.

2 Related Work

2.1 Class-Incremental Learning

Non-prompt-based approaches. Based on their technical principles, existing CIL approaches can be broadly divided into three major categories. Replay-based methods (Rebuffi et al., 2017; Aljundi et al., 2019; Rolnick et al., 2019; Wang et al., 2024; 2023c) mitigate catastrophic forgetting by replaying a small subset of representative samples from old classes when training on new batches of data. Regularization-based methods (Kirkpatrick et al., 2017; Zenke et al., 2017) preserve old knowledge by imposing regularization constraints that limit the size of parameter updates during training on new data. Parameter-isolation-based methods (Mallya & Lazebnik, 2018; Serra et al., 2018; Li & Hoiem, 2017; Lee et al., 2020) partition the model parameters into separate modules dedicated to old and new classes, respectively.

Prompt-based approaches. Prompt-based methods represent an important paradigm of lightweight fine-tuning for pre-trained models in class-incremental learning (CIL). L2P was the pioneering work that introduced prompt learning into continual learning. It uses a key-query matching mechanism where each prompt in the pool is assigned a learnable key. DualPrompt (Wang et al., 2022b) decomposes prompts into two complementary sets: general prompts and expert prompts. CODA-Prompt (Smith et al., 2023) computes attention weights over all keys using the query vector during prompt retrieval, and constructs the final adaptive prompt via a weighted sum of all prompt components guided by these attention weights. The most recent HideP (Wang et al.) proposes hierarchical prompt

components and optimizes them separately based on empirical analysis. Previous methods inherently rely on sufficient labeled samples to learn and consolidate task-specific knowledge into prompts.

2.2 Few-Shot Class-Incremental Learning

Pioneered by TOPIC (Tao et al., 2020), the standard FSCIL protocol requires a model to first establish a robust foundation on a data-rich set of base classes, and subsequently learn new tasks incrementally from few-shot examples without suffering from catastrophic forgetting. Existing FSCIL methods fundamentally into two primary paradigms (Zhao et al., 2023). The first paradigm, often characterized as forward-compatible learning (Zhu et al., 2021; Ahmed et al., 2024; Song et al., 2023; Wang et al., 2023b; Yang et al., 2023; Zhou et al., 2022b), focuses on cultivating highly generalizable representations during the base training phase. While this ensures stability, it inherently constricts the model's capacity to absorb new semantic concepts. Conversely, backward-compatible approaches (Tao et al., 2020; Shi et al., 2021; Cheraghian et al., 2021; Dong et al., 2021; Zhao et al., 2023; Kim et al., 2023; Mazumder et al., 2021; Zhang et al., 2021b; Zhuang et al., 2023) attempt to incrementally learn new classes by actively updating the network parameters. However, tuning the backbone on scarce incremental data inevitably leads to severe few-shot overfitting and the catastrophic forgetting of base classes. Recently, the integration of ViT and prompt-based learning has emerged as an effective strategy to navigate this dilemma. Rather than updating the entire network, methods like PriViLege (Park et al., 2024) and ASP (Liu et al., 2024) introduce learnable prompt parameters to encod e domain-specific or task-invariant knowledge while keeping the core backbone frozen. Despite these architectural advances, most existing approaches often neglecting the broader contextual dynamics required for knowledge transfer. To address this limitation, our approach starts from the perspective of class relation modeling, leverages the exploitable hierarchical semantic relations among classes as the structured prior, and introduces richer semantic prior knowledge during the incremental phase.

3 Preliminaries

3.1 Problem Formulation

Given a sequence of temporally ordered training tasks $D_0, D_1, ..., D_T$, where each task D_t ($t \in \{0, 1, ..., T\}$) contains a disjoint class set C_t such that $C_i \cap C_j = \emptyset$ for any $i < j$, and the complete class space is $C = \bigcup_{t=0}^{T} C_t$. The first task D_0, referred to as the base task, contains a sufficient number of samples per class and is used to train the model's fundamental feature extraction and classification capabilities. For each subsequent task D_t ($t \geq 1$), referred to as the incremental task, each class has an extremely limited number of samples satisfying the few-shot constraint (typically k samples per class, $k \in \{1, 5, 10\}$, denoted as k-shot), and no samples from previous classes are included. The objective of FSCIL is to construct a model that can be trained incrementally on successive tasks of data. After training on the t-th task, the model should not only accurately recognize the new classes C_t under the few-shot constraint but also preserve its recognition performance on the old classes from the first $t-1$ tasks $\bigcup_{i=0}^{t-1} C_i$.

3.2 Prompt-based approaches

The essence of a prompt module is a set of learnable parameters $P \in \mathbb{R}^{L_p \times D \times M}$, where L_p denotes the prompt length, D is the dimension of pre-trained representations, and M denotes the number of prompts. These parameters are prepended to the input features x_e of the pre-trained model, such that the model performs inference with $[P, x_e]$ as input. By minimizing the cross-entropy loss to optimize both the prompt parameters and the classification head, we achieve the encoding of task-specific information:

$$\min_{P,W} \sum_{(x,y)\in\mathcal{D}^b} \ell\left(W^\top \phi(x; P), y\right) \tag{1}$$

where $\mathcal{D}^b = \{\mathcal{X}_b, \mathcal{Y}_b\}$ denotes the set of input instances and their corresponding labels of task b, i.e., $\mathcal{X}_b = \{x_i\}_{i=1}^{n_b}$ and $\mathcal{Y}_b = \{y_i\}_{i=1}^{n_b}$. $\phi(x; P)$ represents the feature extractor augmented with prompts, and $\ell(\cdot, \cdot)$ is the cross-entropy loss function.

Prompt Selection. Given a pool of prompts, a core challenge in prompt-based methods is determining which prompt to use for a specific instance, which motivates the development of prompt retrieval techniques. L2P (Wang et al., 2022c) and the E-Prompt component in DualPrompt adopt a key-query matching strategy, where each prompt is assigned a learnable key $\mathbf{k} \in \mathbb{R}^{D\times M}$. The prompt pool is defined as $\mathbf{P} = \{(\mathbf{k}_1, P_1), (\mathbf{k}_2, P_2), \cdots, (\mathbf{k}_M, P_M)\}$, where M is the total number of prompts, and the set of all keys is $\mathbf{K} = \{k_i\}_{i=1}^{M}$. The query vector $\phi(x)$ matches against the prompt pool using the cosine similarity function $\gamma(\cdot,\cdot)$,and the top-N most similar keys $\mathbf{K}_x$ are selected by optimizing the following objective:

$$\mathbf{K}_x = \underset{\{s_i\}_{i=1}^{N}\subseteq[1,M]}{\arg\max} \sum_{i=1}^{N} \gamma(\phi(x), k_{s_i}). \tag{2}$$

Prompt Combination. CODA-Prompt (Smith et al., 2023) advances prompt retrieval by computing attention scores over all keys using the query vector $\phi(x)$, and then constructs a combined prompt via weighted summation of the prompt components. The prompt combination formulas are as follows:

$$P = \sum_{m=1}^{M} \alpha^m P^m \tag{3}$$

$$\alpha^m = \gamma(\phi(x) \odot \mathbf{A}^m, \mathbf{k}^m) \tag{4}$$

where $\mathbf{A}_m \in \mathbb{R}^{D\times M}$ is a learnable attention vector associated with each prompt, and $\odot$ denotes the Hadamard product. Eq. 4 computes the attention scores between input features and prompt keys via element-wise multiplication, which serve as weights for prompt aggregation. Consequently, prompts whose keys are more similar to the query example contribute more significantly to the final combined prompt. Building upon this framework, SEC-Prompt (Liu & Yang, 2025) proposes Discriminative Prompt (D-Prompt) and Non-Discriminative Prompt (ND-Prompt). The key difference is that D-Prompt applies the Softmax operation to the similarity scores:

$$\alpha_{\mathrm{D}}^m = \frac{e^{\alpha_{\mathrm{D}}^m}}{\sum_k e^{\alpha_{\mathrm{D}}^m}} \tag{5}$$

where α_{D}^m is the similarity score for the m-th D-Prompt. In contrast, ND-Prompt employs reverse attention as the matching mechanism and uses the ReLU function to truncate negative weights:

$$\alpha_{\mathrm{ND}}^m = \max(0, -\gamma(q(x) \odot A_{\mathrm{ND}}^m, K_{\mathrm{ND}}^m)) \tag{6}$$

4 Proposed Method

To address the limitations of existing prompt-based FSCIL approaches that overlook hierarchical semantic relationships among classes, we propose a novel Superclass-guided Hierarchical Prompt learning framework (SH-Prompt), as illustrated in Fig. 1. Semantically related classes often share common visual-semantic patterns, which can be learned collectively rather than independently for each class. SH-Prompt therefore organizes related classes into semantic superclasses and learns shared superclass-level prompts to encode these common patterns. Specifically, we first construct a semantic superclass structure from the base classes using an LLM and assign newly introduced classes to the established superclasses during incremental learning. The resulting superclass prompts are shared by classes within the same superclass, allowing related base classes to learn common visual-semantic patterns and enabling newly introduced classes to reuse the corresponding superclass prompts. We further employ a superclass-guided semantic regularization objective to improve the alignment and discriminability of representations across different superclasses.

4.1 Superclass Partition via LLM

To establish a shared semantic structure across different tasks, we employ an LLM to organize fine-grained classes into semantic superclasses, thereby constructing a superclass-subclass hierarchy. This structure groups semantically related classes that may share common visual-semantic patterns, providing the basis for superclass-level prompt learning. The superclass structure is constructed from the base classes and progressively used to assign newly introduced classes during incremental learning.

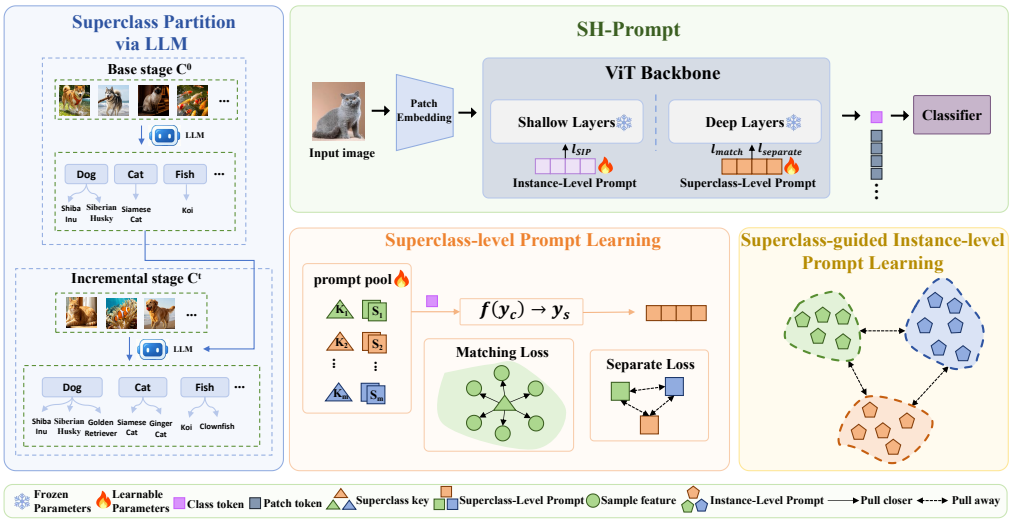

Figure 1: The Overall Training Pipeline of SH-Prompt. First, image labels are organized into a hierarchical superclass–subclass structure via LLM, and the corresponding mapping function f is constructed. The Superclass-level Prompt Learning retrieves superclass-level prompts from the prompt pool according to f , and injects them into deep Transformer layers, optimized by $\mathcal{L}_{\text{match}}$ and $\mathcal{L}_{\text{separate}}$. Meanwhile, instance-level prompts generated by SEC-Prompt are injected into shallow layers and further optimized by the superclass-guided instance-level prompt learning loss $\mathcal{L}_{\text{SIP}}$ to enhance inter-class separability.

Base-stage Superclass Construction. Given the base class set C_0, the LLM organizes the classes into a set of semantic superclasses $S = \{s_1, s_2, \ldots, s_k\}$, where $k = |S|$. Each superclass is represented by a name n_j and a visual semantic profile $p_j : s_j = (n_j, p_j)$. The name provides a concise semantic abstraction, while the visual semantic profile describes the shared visual characteristics of the corresponding classes. Each base class is assigned to one superclass, yielding the base-stage mapping $f_0 : C_0 \to S$.

Incremental-stage Superclass Assignment. When an incremental task D_t arrives, the newly introduced classes C_t are assigned to the established superclass set S before training on the current task. The LLM determines the assignment by considering the semantic information of each new class together with the name and visual semantic profile of the established superclasses. Each new class is assigned to the most semantically compatible superclass within the established semantic space, while the superclass set S remains fixed throughout incremental learning. The superclass-subclass semantic mapping is therefore progressively extended over the task sequence, yielding the overall mapping $f : C \to S$. For each class label y_c, its corresponding superclass label is given by $y_s = f(y_c)$, which provides the superclass-level semantic supervision for subsequent superclass prompt learning.

4.2 Superclass-level Prompt Learning

We propose a superclass semantic-guided prompt learning method to capture shared patterns among classes at a higher semantic level. Specifically, given a predefined superclass partition with k superclasses, we construct a superclass prompt pool $Q_s = \{(S_i^{(l)}, K_i) \mid i = 1, \ldots, k,\ l \in \mathcal{L}\}$, where $S_i^{(l)} \in \mathbb{R}^{L_e \times D}$ denotes the prompt parameters at the l-th layer, and $K_i \in \mathbb{R}^D$ denotes the corresponding key. Here, L_e is the prompt length, D is the feature dimension of the backbone network, and $\mathcal{L}$ represents the set of Transformer layers where prompts are injected. To enhance the diversity and stability among different superclass prompts, we apply Gram-Schmidt orthogonalization to initialize both the prompt parameters $S_i^{(l)}$ and key vectors K_i. Given an input sample x, we extract its global feature $x_{cls} \in \mathbb{R}^D$ from the [CLS] token of the final Transformer layer in ViT and use it as a query to compute the cosine similarity with each superclass key. During training,

we use a label-supervised prompt selection strategy. For each sample (x, y_c), we first obtain its superclass label $y_s = f(y_c)$ and directly select the corresponding prompt $S_{y_s}^{(l)}$ to adapt the feature representations. Meanwhile, inspired by DualPrompt (Wang et al., 2022b), we incorporate a matching mechanism to enhance the consistency between sample representations and their corresponding superclass semantics. Specifically, we define a matching loss that explicitly constrains the similarity between the sample feature x_{cls} and its corresponding superclass key K_{y_s} in the feature space, pulling them closer while keeping them distinct from other superclass semantics.

$$\mathcal{L}_{\text{match}} = \frac{1}{N} \sum_{n=1}^{N} \left(1 - \gamma(x_{cls}^{(n)}, K_{y_s^{(n)}})\right) \tag{7}$$

where N is the batch size. Furthermore, to enhance discriminability, we introduce a separation regularization that discourages different superclasses from being overly similar in the feature space. Specifically, we minimize the similarity between superclass representations within a batch, thereby encouraging stronger separation between different superclasses. It is defined as follows:

$$\mathcal{L}_{\text{separate}} = \frac{1}{|\mathcal{B}_s|} \sum_{i \neq j} \gamma(S_i, S_j) \tag{8}$$

where S_i and S_j denote the superclass prompt representations of the i-th and j-th superclasses, and $\mathcal{B}_s$ denotes the set of superclasses present in the current batch. During inference, instead of relying on label information, we perform similarity-based prompt retrieval. Specifically, we select $i^* = \arg\max_i \gamma(x_{cls}, K_i)$, and use the corresponding prompt $S_{i^*}^{(l)}$ to modulate the feature representation. In both training and inference stages, we employ Prefix Tuning (Pre-T) (Wang et al., 2022b)to inject prompts into the Transformer. We combine the above objectives to form the super-class semantic regularization loss:

$$\mathcal{L}_{\text{SP}} = \mathcal{L}_{\text{match}} + \mathcal{L}_{\text{separate}}, \tag{9}$$

4.3 Superclass-guided Instance-level Prompt Learning

To fully exploit the hierarchical semantic structure of the superclass-subclass hierarchy, we introduce a superclass-guided instance-level prompt learning loss based on the D-Prompt module of SEC-Prompt. Specifically, we design an inter-superclass loss as the optimization objective from the perspective of constraining the feature distribution of prompts. First, we construct a symmetric mask matrix based on the superclass labels of samples within a batch for sample pair filtering in the loss function. The cross-superclass mask matrix M_{cross} is of size (B, B), where $M_{\text{cross}}[i, j] = 1$ if sample i and sample j belong to different superclasses, and 0 otherwise. Based on the above mask matrix, the superclass-guided instance-level prompt learning loss is defined to enhance discriminability by minimizing the similarity between image-level prompt features of samples from different superclasses:

$$\mathcal{L}_{\text{SIP}} = \frac{1}{N_{\text{cross}}} \sum_{i=1}^{B} \sum_{j=1}^{B} \gamma(P_i, P_j) \cdot M_{\text{cross}}[i, j] \tag{10}$$

where N_{cross} denotes the total number of cross-superclass sample pairs within the batch. By explicitly reducing the similarity between samples from different superclasses, this loss encourages more discriminative feature representations across distinct semantic groups and helps reduce feature confusion. Finally, the overall optimization objective of the model is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L} + \lambda_1 \cdot \mathcal{L}_{\text{SP}} + \lambda_2 \cdot \mathcal{L}_{\text{SIP}} \tag{11}$$

where $\mathcal{L}$ is the loss function to ensure discriminability of sample features and instance-level prompts, λ_1 and λ_2 are the weight coefficients of the corresponding loss terms. We employ the cross-entropy loss and the discriminative prompt clustering loss in the SEC-Prompt method to compute $\mathcal{L}$.

5 Experiments

5.1 Experimental Setups

Datasets. In this chapter, we follow the same dataset splits and task protocols as SEC-Prompt to ensure fair and consistent comparisons with existing methods. CIFAR-100 (Krizhevsky et al.,

Table 1: Results (%) on CIFAR-100 dataset.

Method	Acc. in each session ↑									A_{avg} ↑	PD ↓	HAcc ↑
	0	1	2	3	4	5	6	7	8			
CEC(Zhang et al., 2021a)	91.6	88.1	85.3	81.7	80.2	78.0	76.5	74.8	72.6	81.0	19.0	64.1
FACT(Zhou et al., 2022a)	91.0	87.2	83.5	79.7	77.2	74.8	73.1	71.6	69.4	78.6	21.7	55.5
L2P(Wang et al., 2022c)	84.7	82.3	80.1	77.5	77.0	76.0	75.6	74.1	72.3	77.7	12.4	68.0
DualPrompt(Wang et al., 2022b)	86.0	83.6	82.9	80.2	80.6	80.2	80.5	79.0	77.4	81.1	8.5	75.3
TEEN(Wang et al., 2023a)	**92.9**	90.2	88.4	86.8	86.4	86.0	85.8	85.1	84.0	87.3	8.8	81.2
CODA-Prompt(Smith et al., 2023)	86.0	83.6	81.6	79.2	79.1	78.5	78.3	77.0	75.4	79.9	10.6	72.2
PriViLege(Park et al., 2024)	90.9	89.4	89.0	87.6	87.9	87.4	87.6	87.2	86.1	88.1	4.8	84.7
ASP(Liu et al., 2024)	92.2	**90.7**	90.0	88.7	88.7	88.2	88.2	87.8	86.7	89.0	5.5	85.3
LGSP(Jiang et al., 2025)	86.9	83.8	84.2	81.9	82.3	80.9	81.0	80.1	78.2	84.9	8.7	82.2
SEC-Prompt(Liu & Yang, 2025)	91.8	90.5	**90.4**	88.9	89.1	89.0	88.9	86.9	83.4	88.8	8.4	83.4
Ours	91.8	90.5	**90.4**	**90.0**	**89.2**	**90.0**	**89.1**	**88.8**	**87.6**	**89.5**	**4.2**	**86.9**

Table 2: Results (%) on ImageNet-R dataset.

Method	Acc. in each session ↑											A_{avg} ↑	PD ↓	HAcc ↑
	0	1	2	3	4	5	6	7	8	9	10			
CEC(Zhang et al., 2021a)	79.4	71.9	69.0	64.1	60.4	58.6	56.4	53.2	52.0	50.0	48.3	60.3	31.1	32.6
FACT(Zhou et al., 2022a)	79.4	72.5	69.0	63.8	60.1	57.6	54.7	52.2	50.2	48.1	46.0	59.4	33.4	22.3
L2P(Wang et al., 2022c)	73.9	70.9	69.3	65.9	64.0	62.6	60.1	59.5	59.0	58.2	56.8	63.7	17.2	52.0
DualPrompt(Wang et al., 2022b)	71.5	69.0	69.0	67.4	66.6	65.9	64.1	64.0	64.0	63.4	62.5	66.1	**9.0**	61.6
TEEN(Wang et al., 2023a)	**84.6**	76.7	68.8	67.6	64.3	60.6	58.3	56.1	56.1	54.7	54.9	63.9	29.7	45.4
CODA-Prompt(Smith et al., 2023)	82.8	80.1	78.6	75.3	73.1	71.8	70.0	69.2	68.8	67.4	66.4	73.0	16.8	61.7
PriViLege(Park et al., 2024)	80.0	77.3	77.0	73.8	72.8	72.3	71.3	69.7	69.5	67.7	66.5	72.5	13.5	64.7
ASP(Liu et al., 2024)	83.3	80.4	79.6	77.0	75.6	74.7	73.0	72.1	71.9	70.9	69.7	75.3	13.5	67.0
SEC-Prompt(Liu & Yang, 2025)	82.8	80.3	79.7	77.3	76.5	75.5	74.1	73.7	73.6	73.1	72.3	76.3	10.5	71.3
Ours	83.3	**81.0**	**80.7**	**78.5**	**79.1**	**77.4**	**76.2**	**75.8**	**75.5**	**74.9**	**73.7**	**77.7**	9.6	**73.3**

2009): We designate 60 classes as base classes, and split the remaining 40 classes into 8 incremental sessions, each following a 5-way 5-shot setting. ImageNet-R (Hendrycks et al., 2021): We designate 100 classes as base classes, and split the remaining 100 classes into 10 incremental sessions, each following a 10-way 5-shot setting. CUB200 (Wah et al., 2011): We designate 100 classes as base classes, and split the remaining 100 classes into 10 incremental sessions, each following a 10-way 5-shot setting.

Baselines. We compare our method against representative prompt-based CIL methods, including L2P (Wang et al., 2022c), DualPrompt (Wang et al., 2022b), and CODA-Prompt (Smith et al., 2023). Furthermore, We also benchmark against representative FSCIL algorithms, including non-prompting methods CEC (Zhang et al., 2021a), FACT (Zhou et al., 2022a), and TEEN (Wang et al., 2023a), and prompt-based methods PriViLege (Park et al., 2024), ASP (Liu et al., 2024), and SEC-Prompt (Liu & Yang, 2025). See Appendix for more details of experimental setups.

5.2 Comparison Results

This section reports the performance of baseline methods and our SH-Prompt under the FSCIL setting, with detailed results summarized in Tab. 1, Tab. 2 and Tab. 3. Overall, SH-Prompt achieves outstanding performance across all three benchmark datasets. In particular, it reaches SOTA performance on both A_{avg} and HAcc, while maintaining a low level of PD. Specifically, on CIFAR-100, SH-Prompt outperforms the previous best method by 3.5%, while also achieving the lowest PD, demonstrating its effectiveness in preserving previously learned knowledge while adapting to novel categories. On ImageNet-R, its A_{avg} and HAcc surpass the second-best method by 1.4% and 2.0%, respectively, further demonstrating the effectiveness of superclass-level knowledge sharing on a dataset with diverse visual categories. On CUB-200, although the performance improvement is relatively smaller, its A_{avg} still exceeds the previous best method by 0.4%. This may be because all categories in CUB-200 belong to the bird category and already share substantial visual and semantic characteristics. Therefore, further grouping them into semantic superclasses may provide relatively limited additional benefits. These consistent results demonstrate that SH-Prompt can effectively balance discriminability and generalization capability, delivering superior incremental learning performance on benchmarks with varying difficulty levels and data distribution characteristics.

Table 3: Results (%) on CUB200 dataset.

Method	Acc. in each session ↑											A_{avg} ↑	PD ↓	HAcc ↑
	0	1	2	3	4	5	6	7	8	9	10			
CEC(Zhang et al., 2021a)	84.8	82.5	81.4	78.5	79.3	77.8	77.4	77.6	77.2	76.9	76.8	79.1	7.9	76.2
FACT(Zhou et al., 2022a)	87.3	84.2	82.1	78.1	78.4	76.3	75.4	75.5	74.4	74.1	73.9	78.2	13.4	72.0
L2P(Wang et al., 2022c)	82.4	81.2	79.0	76.8	76.2	74.7	74.1	74.1	72.7	73.0	73.6	76.2	8.7	73.6
DualPrompt(Wang et al., 2022b)	83.5	82.2	80.9	79.5	78.6	77.0	76.3	77.0	75.7	76.1	76.5	78.5	7.0	76.3
TEEN(Wang et al., 2023a)	**89.0**	86.5	85.9	83.3	83.3	82.2	82.1	80.4	80.5	80.1	80.6	83.1	9.4	80.2
CODA-Prompt(Smith et al., 2023)	79.6	78.1	76.4	75.6	75.0	73.1	72.5	72.8	72.0	72.4	72.9	74.6	6.7	72.5
PriViLege(Park et al., 2024)	82.3	81.3	80.5	77.8	77.8	76.0	75.7	76.0	75.2	75.2	75.1	77.5	7.2	72.3
ASP(Liu et al., 2024)	87.1	86.0	84.9	83.4	83.6	82.4	82.6	83.0	82.6	83.0	83.5	83.8	3.6	83.4
LGSP(Jiang et al., 2025)	83.7	82.6	80.8	78.1	77.6	76.0	75.7	76.2	75.4	75.6	75.7	77.9	8.0	77.6
SEC-Prompt(Liu & Yang, 2025)	87.2	86.5	85.4	84.5	84.5	83.4	83.3	**83.9**	**83.5**	**83.7**	**84.0**	84.5	**3.3**	83.9
Ours	87.8	**87.0**	**85.9**	**85.0**	**85.3**	**84.1**	**83.8**	83.7	83.3	83.6	**84.0**	**84.9**	3.8	**84.0**

Table 4: Ablation study on ImageNet-R dataset for FSCIL. SP denote Superclass Prompt.

SP	$\mathcal{L}_{\text{SP}}$	$\mathcal{L}_{\text{SIP}}$	Acc. in each session ↑											A_{avg} ↑	PD ↓	HAcc ↑
			0	1	2	3	4	5	6	7	8	9	10			
			82.8	80.3	79.7	77.3	76.5	75.5	74.1	73.7	73.6	73.1	72.3	76.3	10.5	71.3
✓			**83.3**	**81.0**	80.5	78.2	77.6	77.0	75.7	75.0	74.9	74.3	73.3	77.3	10.0	72.8
✓	✓		**83.3**	**81.0**	80.5	78.2	77.7	77.1	75.9	75.6	**75.5**	74.6	73.6	77.5	**9.6**	73.2
✓	✓	✓	**83.3**	**81.0**	**80.7**	**78.5**	**79.1**	**77.4**	**76.2**	**75.8**	**75.5**	**74.9**	**73.7**	**77.7**	**9.6**	**73.3**

5.3 Ablation Study

We conduct extensive ablation studies to validate the effectiveness of each proposed component in our framework. We take SEC-Prompt as the backbone baseline on the ImageNet-R dataset, and incrementally integrate each of our proposed modules into the baseline for evaluation. As shown in Tab. 4, we observe that introducing the Superclass Prompt into the SEC-Prompt baseline yields a significant performance gain, which verifies that the Superclass Prompt injects complementary high-level semantic information into the baseline framework. Subsequently, incorporating the proposed $\mathcal{L}_{\text{SP}}$ loss term brings an improvement in HAcc, demonstrating that this loss function effectively achieves a better trade-off between stability and plasticity. Finally, with the introduction of the $\mathcal{L}_{\text{SIP}}$ regularization term, the model achieves consistently superior accuracy on subsequent incremental tasks and further improves the overall performance. This result highlights the critical role of $\mathcal{L}_{\text{SIP}}$ in preserving discriminative feature representations and mitigating catastrophic forgetting throughout the incremental learning phases.

5.4 Further Analysis

Hyper-Parameter. We conduct a thorough sensitivity analysis on the hyperparameters λ_1 and λ_2 across all three benchmark datasets. Specifically, λ_1 is introduced to tune the weight of the loss term associated with the Superclass Prompt, while λ_2 governs the weight of the superclass-guided instance-level prompt learning loss. The corresponding results are shown in Fig. 2. Across the comprehensive experimental results on all datasets, the model achieves the optimal overall performance in terms of A_{avg} when $\lambda_1 = \lambda_2 = 0.5$.

Effectiveness of Superclass-level Prompt Learning. To investigate whether the performance gains mainly arise from the proposed Superclass-level Prompt Learning mechanism rather than the additional semantic supervision provided by the LLM, we compare our LLM-guided partition with an automatic prototype-guided partition. The latter constructs superclasses solely from visual class prototypes without using LLM-generated semantic relationships. On ImageNet-R, the prototype-guided strategy generates 19 superclasses. As shown in Tab. 5, the prototype-guided strategy already outperforms the baseline, demonstrating that the proposed superclass prompt learning mechanism itself can provide substantial performance gains without relying on LLM-generated semantic supervision. Furthermore, the LLM-guided partition achieves better performance than the prototype-guided strategy, indicating that incorporating semantic relationships among classes can further improve the quality of the superclass partition.

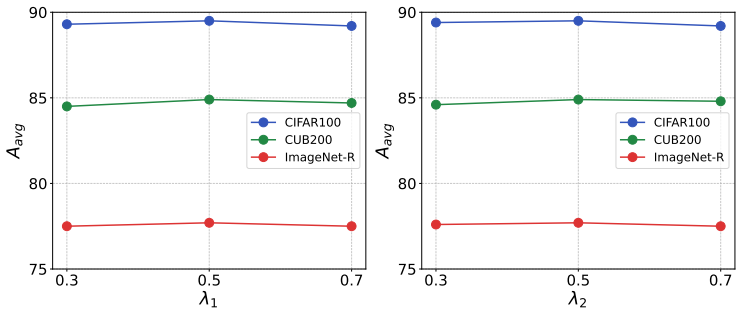

Figure 2: Hyperparameter Analysis.

Table 5: Effectiveness of Superclass-lever Prompt Learning on ImageNet-R.

Method	Acc. in each session ↑											A_{avg} ↑	PD ↓	HAcc ↑
	0	1	2	3	4	5	6	7	8	9	10			
SEC-Prompt(Liu & Yang, 2025)	82.8	80.3	79.7	77.3	76.5	75.5	74.1	73.7	73.6	73.1	72.3	76.3	10.5	71.3
Prototype-guided	83.0	80.7	80.0	78.1	77.5	77.1	76.1	75.6	**75.5**	74.6	**73.7**	77.4	**9.2**	**73.3**
LLM-guided	83.3	**81.0**	**80.7**	**78.5**	**78.1**	**77.4**	**76.2**	**75.8**	**75.5**	**74.9**	**73.7**	**77.7**	9.6	**73.3**

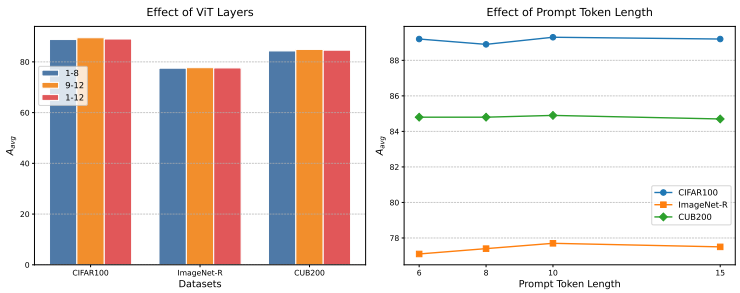

Figure 3: Superclass Prompt Analysis.

Superclass Prompt Analysis. We analyze the design of the superclass Prompt from two key aspects: the insertion layer positions and the length of prompt tokens, as illustrated in Fig. 3.In terms of insertion layers, the D-Prompt is fixedly inserted into layers 1–8 following the original protocol, while we insert the superclass Prompt into layers 1–8, 1–12, and 9–12 respectively for ablation study. The results demonstrate that inserting the superclass Prompt into the deep layers (9–12) achieves the optimal performance. In contrast, inserting it into shallow layers or the full network layers leads to a consistent performance drop. This phenomenon is mainly attributed to the complementary functionalities of the two types of Prompts across different semantic levels: D-Prompt focuses on capturing instance-level discriminative features in shallow layers, while the superclass Prompt is more suitable for modeling cross-category semantic relations in deep layers. In terms of Prompt length, we further conduct ablation studies by setting the length of the superclass Prompt to 6, 8, 10, and 15, respectively. Experimental results show that the model achieves the best performance when the Prompt length is set to 10, while either insufficient or excessive tokens will result in performance degradation.

6 Conclusion

We propose SH-Prompt, a superclass-guided hierarchical prompt learning framework for FSCIL that exploits hierarchical semantic relationships for cross-class knowledge sharing. By constructing a superclass–subclass hierarchy with LLM-based superclass partitioning, we introduce a key-based superclass prompt pool and a superclass-guided instance-level prompt learning loss to promote semantic knowledge sharing and feature discriminability. Experiments on multiple benchmarks show that our method outperforms existing approaches, demonstrating the effectiveness of hierarchical semantic information.

7 FINAL INSTRUCTIONS

AI USE STATEMENT

We used generative AI tools only for literature search, code debugging, and language editing. We did not use them for research ideas, results, or other core contributions. All AI-assisted outputs were reviewed: references were cross-checked against original sources, code was tested, and text was revised. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve human subjects, sensitive personal data, or other issues requiring ethical consideration.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of our experimental settings to facilitate reproducibility. The datasets used in our experiments are described in 5.1, while the training details, including the main hyperparameters and optimization settings, are provided in Appendix B. The complete prompts used for LLM-based superclass partition and incremental superclass assignment are provided in Appendix A. The main paper and Appendix together provide the methodological, experimental, and implementation details necessary to reproduce our results.

REFERENCES

Noor Ahmed, Anna Kukleva, and Bernt Schiele. Orco: Towards better generalization via orthogonality and contrast for few-shot class-incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 28762–28771, 2024.

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ali Cheraghian, Shafin Rahman, Pengfei Fang, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Semantic-aware knowledge distillation for few-shot class-incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2534–2543, 2021.

Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. Metafscil: A meta-learning approach for few-shot class incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14166–14175, 2022.

Songlin Dong, Xiaopeng Hong, Xiaoyu Tao, Xinyuan Chang, Xing Wei, and Yihong Gong. Few-shot class-incremental learning via relation knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1255–1263, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021.

Robert M French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3 (4):128–135, 1999.

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 *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8340–8349, 2021.

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

Do-Yeon Kim, Dong-Jun Han, Jun Seo, and Jaekyun Moon. Warping the space: Weight space rotation for class-incremental few-shot learning. In *Proceedings of the International Conference on Learning Representations*, 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. 114(13):3521–3526, 2017.

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

Anna Kukleva, Hilde Kuehne, and Bernt Schiele. Generalized and incremental few-shot learning by explicit learning and calibration without forgetting. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9020–9029, 2021.

Janghyeon Lee, Donggyu Joo, Hyeong Gwon Hong, and Junmo Kim. Residual continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4553–4560, 2020.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017.

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 *Proceedings of the Europeon Conference on Computer Vision*, pp. 1–18, 2024.

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

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7765–7773, 2018.

Pratik Mazumder, Pravendra Singh, and Piyush Rai. Few-shot lifelong learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2337–2345, 2021.

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 Conference on Computer Vision and Pattern Recognition*, pp. 23881–23890, 2024.

Can Peng, Kun Zhao, Tianren Wang, Meng Li, and Brian C Lovell. Few-shot class-incremental learning from an open-set perspective. In *Proceedings of the Europeon Conference on Computer Vision*, pp. 382–397, 2022.

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*, pp. 2001–2010, 2017.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *Proceedings of the International Conference on Machine Learning*, pp. 4548–4557, 2018.

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.

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 Conference on Computer Vision and Pattern Recognition*, pp. 11909–11919, June 2023.

Zeyin Song, Yifan Zhao, Yujun Shi, Peixi Peng, Li Yuan, and Yonghong Tian. Learning with fantasy: Semantic-aware virtual contrastive constraint for few-shot class-incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 24183–24192, 2023.

Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12183–12192, 2020.

Anh-Khoa Nguyen Vu, Thanh-Toan Do, Nhat-Duy Nguyen, Vinh-Tiep Nguyen, Thanh Duc Ngo, and Tam V Nguyen. Instance-level few-shot learning with class hierarchy mining. *IEEE Transactions on Image Processing*, 32:2374–2385, 2023.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Liyuan Wang, Jingyi Xie, Xingxing Zhang, Mingyi Huang, Hang Su, and Jun Zhu. Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, pp. 69054–69076. Curran Associates, Inc.

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. *arXiv preprint arXiv:2312.05229*, 2023a.

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. *Advances in Neural Information Processing Systems*, pp. 15060–15076, 2023b.

Zhenyi Wang, Li Shen, Le Fang, Qiuling Suo, Donglin Zhan, Tiehang Duan, and Mingchen Gao. Meta-learning with less forgetting on large-scale non-stationary task distributions. In *Proceedings of the Europeon Conference on Computer Vision*, pp. 221–238, 2022a.

Zhenyi Wang, Li Shen, Tiehang Duan, Qiuling Suo, Le Fang, Wei Liu, and Mingchen Gao. Distributionally robust memory evolution with generalized divergence for continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):14337–14352, 2023c.

Zhenyi Wang, Yan Li, Li Shen, and Heng Huang. A unified and general framework for continual learning. In *The Twelfth International Conference on Learning Representations*, 2024.

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 *Proceedings of the Europeon Conference on Computer Vision*, pp. 631–648, 2022b.

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 Conference on Computer Vision and Pattern Recognition*, pp. 139–149, 2022c.

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 *Proceedings of the International Conference on Learning Representations*, 2023.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the International Conference on Machine Learning*, pp. 3987–3995, 2017.

Chen Zhang, Nan Song, Guosheng Lin, Yun Zheng, Ping Pan, and Ying Xu. Few-shot incremental learning with continually evolved classifiers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12455–12464, 2021a.

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 Conference on Computer Vision and Pattern Recognition*, pp. 12455–12464, 2021b.

Linglan Zhao, Jing Lu, Yunlu Xu, Zhanzhan Cheng, Dashan Guo, Yi Niu, and Xiangzhong Fang. Few-shot class-incremental learning via class-aware bilateral distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11838–11847, 2023.

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

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 Conference on Computer Vision and Pattern Recognition*, pp. 9046–9056, 2022b.

Kai Zhu, Yang Cao, Wei Zhai, Jie Cheng, and Zheng-Jun Zha. Self-promoted prototype refinement for few-shot class-incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6801–6810, 2021.

Huiping Zhuang, Zhenyu Weng, Run He, Zhiping Lin, and Ziqian Zeng. Gkeal: Gaussian kernel embedded analytic learning for few-shot class incremental task. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7746–7755, 2023.

A LLM PROMPTS

A.1 SUPERCLASS PARTITION PROMPT.

We adopt GPT-5.6 as the LLM in our method and use the following prompt to construct the semantic superclass partition. The prompt emphasizes visual similarity while incorporating semantic knowledge and commonsense relationships.

Role: You are an expert in multimodal visual concept abstraction. Your task is to cluster a set of fine-grained categories into several superclasses based on visual characteristics, semantic knowledge, and commonsense, and to provide a concise name for each superclass.
Important Rules:

1. Category IDs are identifiers only and do not indicate order, distance, or similarity. Do not group categories based on their IDs.
2. Prioritize grouping criteria in the following order: visual similarity $>$ biological, functional, or scene-level semantics $>$ other commonsense relationships.
3. Choose the most direct and informative superclass concept and avoid overly general categories. For example, golden retriever, Labrador, husky, and beagle should be grouped as dog rather than animal.
4. By default, each fine-grained category should belong to exactly one superclass.
5. Superclass names should be concise, general, and easy to understand.
6. For each superclass, provide its member category IDs, a superclass name, and a concise semantic description that captures the shared characteristics of its members. The description should be sufficiently general to support superclass assignment of newly introduced categories in later incremental sessions.
7. All category IDs start from 0.

A.2 INCREMENTAL SUPERCLASS ASSIGNMENT PROMPT.

For each incremental session, we use the following prompt to assign newly introduced fine-grained categories to the existing superclass set. The assignment is based on the semantic profiles established during the base session.

Role: You are an expert in multimodal visual concept abstraction and incremental hierarchical assignment. Your task is to assign each newly introduced fine-grained category to the most appropriate existing superclass based on its semantic profile, name, and member categories.
Important Rules:

1. The assignment criteria should be prioritized as follows: visual similarity $>$ biological, functional, scene-level, or commonsense semantics $>$ superficial similarity in names or keywords.
2. Prioritize the Semantic Profile of each existing superclass for holistic semantic matching. Do not rely on simple keyword or name matching. Instead, consider the overall semantic concept and the visual and functional characteristics described in each superclass profile.
3. If a superclass does not have a Semantic Profile, use its Name and Members as auxiliary information for the assignment.
4. Each new fine-grained category must be assigned to exactly one existing superclass. Do not create new superclasses.
5. All superclass IDs start from 0 and must preserve the IDs provided in the input.

B EXPERIMENTAL SETUPS

Evaluation protocol. Following prior work, we adopt three standard evaluation metrics for FSCIL tasks to comprehensively assess the performance of our model: Average Accuracy (A_{avg}): After completing all incremental sessions, we denote the Top-1 accuracy on task t as A_t. The average accuracy is computed as $A_{avg} = \frac{\sum_{t=0}^{T} A_t}{T+1}$,which measures the overall classification performance of

the model. Harmonic Accuracy (HAcc) (Peng et al., 2022): The harmonic mean of the base class accuracy and the average novel class accuracy, defined as $A_h = \frac{2 \cdot A_o \cdot A_n}{A_o + A_n}$, where A_o denotes the final base class accuracy and A_n denotes the average classification accuracy across all incremental novel classes. This metric reflects the performance balance between old and new classes. Performance Dropping rate (PD): The difference between the base class accuracy immediately after base training and the base class accuracy after all incremental sessions, defined as $PD = A_0 - A_T$, which quantifies the degree of catastrophic forgetting; a smaller value indicates stronger resistance to forgetting.

Training Details. We adopt ViT-B/16 (Dosovitskiy et al., 2021) pre-trained on the ImageNet-1K dataset (Russakovsky et al., 2015) as our backbone network. Across all datasets, input images are resized to 224×224 pixels, and we use the SGD optimizer for training. The model is trained for 30 epochs during the base session and 10 epochs for each incremental session. During the base session, we set the initial learning rate to 0.1, with a batch size of 64 for ImageNet-R and 32 for both CIFAR-100 and CUB-200. For all incremental sessions, a unified batch size of 16 is used, while the learning rate is set to 0.0005 for ImageNet-R and 0.001 for CIFAR-100 and CUB-200 respectively. We assign a pair of D-Prompt and ND-Prompt components to each individual class. The prompt token lengths $L_D = L_{ND}$ are set to 3 for CIFAR-100 and CUB-200, and 5 for ImageNet-R. The size of the prompt pool is set to half the total number of classes. The superclass prompt token length L_{SP} is set to 10, and the size of the superclass prompt pool is determined by the number of superclasses. The hyperparameters are set as follows:$\lambda_1 = \lambda_2 = 0.5$.

Experimental Environment. All experiments are conducted on an Ubuntu 22.04.1 LTS system using Python 3.11.4 and PyTorch 2.1.2. The proposed method is implemented on a single NVIDIA GeForce RTX 3090 GPU (24GB memory) .

C Automatic Class Prototype-guided Superclass Partition

We design an automatic class prototype-guided superclass partition scheme, with the implementation pipeline detailed as follows: in the base phase, we extract class-specific features from all accessible training samples and compute the corresponding class prototype for each base class; we then obtain the initial superclasses via a clustering algorithm with automatic cluster number determination. For each novel class in the incremental phases, we compute its class prototype from the limited few-shot samples, and assign it to the most similar pre-existing initial superclass based on feature similarity. The core rationale behind this design is that base classes have sufficient training data to generate reliable and representative class prototypes, which serve as a robust baseline for initial superclass partition. In contrast, the class prototypes of novel classes in incremental phases are estimated from extremely limited few-shot samples, leading to insufficient representativeness and high statistical variance. We thus avoid re-clustering for novel classes, and instead assign them to the established initial superclasses via similarity matching.

D Additional Results

D.1 Base & Incremetal Phase Accuracy.

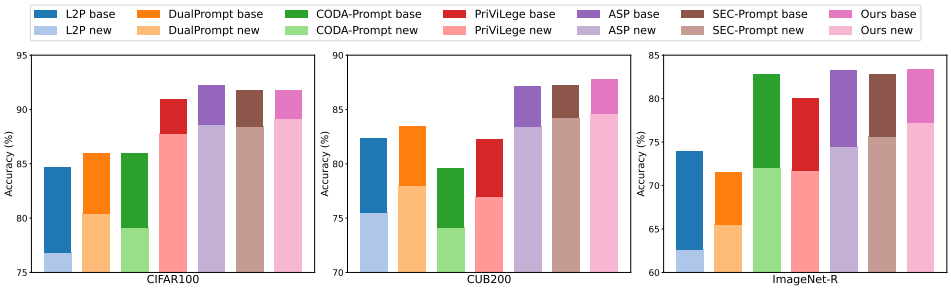

Figure 4: Comparison between base and incremental phase.

Table 6: Robustness to Different Numbers of Superclasses on ImageNet-R.

Method	Acc. in each session ↑											A_{avg} ↑	PD ↓	HAcc ↑
	0	1	2	3	4	5	6	7	8	9	10			
SEC-Prompt(Liu & Yang, 2025)	82.8	80.3	79.7	77.3	76.5	75.5	74.1	73.7	73.6	73.1	72.3	76.3	10.5	71.3
18 Superclasses	83.4	**80.9**	80.2	78.2	77.9	77.3	76.0	75.3	75.2	74.6	**73.7**	77.5	9.7	**73.3**
19 Superclasses	**83.9**	80.7	80.1	78.2	77.6	77.3	76.0	75.4	75.3	74.4	73.2	77.4	9.7	73.0
20 Superclasses	83.3	**81.0**	**80.7**	**78.5**	**78.1**	**77.4**	**76.2**	**75.8**	**75.5**	**74.9**	**73.7**	**77.7**	9.6	**73.3**
21 Superclasses	83.4	**81.0**	80.2	78.2	77.7	77.3	75.9	75.3	75.1	74.4	73.3	77.4	10.0	72.8
22 Superclasses	82.5	80.2	80.1	78.2	77.5	76.7	75.6	75.1	74.8	73.9	73.1	77.1	**9.4**	72.7

Table 7: Robustness to Different LLMs on ImageNet-R.

Method	Acc. in each session ↑											A_{avg} ↑	PD ↓	HAcc ↑
	0	1	2	3	4	5	6	7	8	9	10			
SEC-Prompt(Liu & Yang, 2025)	82.8	80.3	79.7	77.3	76.5	75.5	74.1	73.7	73.6	73.1	72.3	76.3	10.5	71.3
DeepSeek	**83.5**	**81.0**	80.6	78.3	77.4	76.8	75.7	75.2	75.0	74.4	73.5	77.4	10	73.0
Gemini	83.2	**81.0**	80.5	78.2	77.5	76.7	75.6	75.1	75.1	74.3	73.2	77.3	10	72.8
GPT	83.3	**81.0**	**80.7**	**78.5**	**78.1**	**77.4**	**76.2**	**75.8**	**75.5**	**74.9**	**73.7**	**77.7**	**9.6**	**73.3**

In addition to the aforementioned metrics, we further present a comparative analysis of the accuracy in the base phase and the average accuracy across the incremental phases, with the results depicted in Fig. 4. In this figure, the solid-colored regions represent the accuracy achieved in the base class phase, while the corresponding shaded areas denote the average accuracy obtained in the incremental phases. A smaller gap between the solid-colored region and the shaded area within each block indicates a stronger resistance of the model to catastrophic forgetting during the incremental learning process. It can be clearly observed that our method not only preserves robust learning capability on base classes, but also achieves superior ability in learning novel incremental classes.

D.2 Robustness to Superclass Granularity.

We further investigate the robustness of the proposed framework to different superclass granularities. In our LLM-guided strategy, the number of superclasses is automatically determined by the LLM according to the semantic relationships among the classes. To examine the sensitivity of the framework to the partition granularity, we additionally evaluate different numbers of superclasses on ImageNet-R. As shown in Tab. 6, the performance varies only slightly across different numbers of superclasses, while all settings consistently outperform the baseline. These results show that the proposed framework is not highly sensitive to the exact number of superclasses and remains robust across different levels of semantic granularity.

D.3 Robustness to Different LLMs.

To further examine whether the superclass construction depends on a specific LLM, we conduct experiments on ImageNet-R using GPT, Gemini, and DeepSeek to independently construct the superclass structures, while keeping the downstream prompt learning procedure and all other experimental settings unchanged. As shown in Tab. 7, all three LLMs achieve comparable performance and consistently outperform the baseline. These results demonstrate that the effectiveness of SH-Prompt is robust to the choice of LLM and does not rely on a specific LLM for constructing the superclass structure.

E Limitations

Despite its strong performance on FSCIL tasks, the proposed method has several limitations. Its effectiveness depends on the quality of the hierarchical semantic structure, and imprecise superclass partitioning may introduce noise into representation learning. Moreover, while hierarchical semantic sharing facilitates cross-class knowledge transfer, achieving a better balance between semantic sharing and class discriminability is still an open problem.