Pattern Recognition 179 (2026) 113610

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Structured-condensed prompt tuning in vision-language models for fine-grained image recognition

Xinda Liu [a], Qinyu Zhang [a], Weiqing Min [b,*], Guohua Geng [a], Shuqiang Jiang [b]

[a] School of Information Science and Technology, Northwest University, Xi'an, 710127, Shaanxi, China
[b] Laboratory of Intelligent Information Processing, Institute of Computing Technology, Beijing, 100190, Beijing, China

ARTICLE INFO

Keywords:
Prompt tuning
Vision-language models
Semantic relation
Few-shot learning

ABSTRACT

Fine-grained image recognition poses a significant challenge due to the substantial expertise and effort required for manual annotation. Vision-language models (VLMs) like CLIP provide a compelling zero-shot alternative, reducing reliance on extensive labeled data. However, their ability to capture subtle distinctions remains limited, leading to subpar recognition performance. While prompt tuning has proven effective for adapting VLMs, most existing methods treat class labels as isolated, discrete entities, overlooking the rich semantic relationships between them. This oversimplified assumption limits the model's ability to capture hierarchical dependencies and inter-class correlations–both critical for distinguishing visually similar categories. The problem is especially acute in fine-grained classification, where accurate recognition depends on understanding complex label semantics. To address this, we propose Structured-Condensed Prompt Tuning (SCPT), which enhances semantic structure modeling in prompt learning. Specifically, we introduce Semantic Relation Encoding (SRE) to explicitly model inter-class semantic topology and encode structured label relationships. In parallel, we design a Semantic Condensation loss (ScLoss) to suppress redundant supervision and extract discriminative components from the global semantic space. Together, these components significantly improve semantic alignment and fine-grained discrimination. Extensive experiments on 14 fine-grained benchmarks show that SCPT effectively mitigates semantic ambiguity and achieves state-of-the-art performance in both few-shot and base-to-novel generalization settings.

1. Introduction

Fine-grained image recognition (FGIR) plays a pivotal role in scenarios demanding precise differentiation between visually similar subcategories, rich image captioning [1–3], image generation [4,5], food recognition [6–8], and food recommendation [9,10]. The essential challenge stems from the requirement for expert-level annotation precision, which necessitates a nuanced understanding of subtle visual differences, often demanding domain-specific knowledge [11]. The exorbitant cost associated with acquiring such human-expert annotations has emerged as a critical bottleneck, that fundamentally constrains the development of FGIR systems in novel domains.

Vision-language models (VLMs), such as CLIP, offer a promising solution by leveraging large-scale web data for zero-shot learning, thereby mitigating the reliance on extensive manual annotations [12]. CLIP employs contrastive learning to align text and image representations, enabling category classification without task-specific training. While VLMs demonstrate strong generalization in broad categories, this advantage diminishes notably in fine-grained recognition contexts, predominantly stemming from the model's constrained capacity to discern nuanced inter-class variations.

Prompt tuning adapts task-specific prompts to better leverage the latent knowledge in vision-language models (VLMs), driving strong performance on downstream tasks. It falls into two categories: visual and textual prompt tuning, with the latter optimizing text inputs for improved alignment with visual representations. In this work, we focus exclusively on textual prompt tuning, omitting modifications to the visual encoder. Instead of relying on manually crafted prompts, textual prompt tuning replaces static templates with learnable context vectors [13,14], which are dynamically optimized using a small set of training samples to enhance image-text alignment, thereby alleviating data scarcity constraints.

However, existing prompt tuning approaches treat category labels as independent, discrete entities, failing to leverage the rich semantic relationships among them. This oversimplification limits the model's capacity to capture hierarchical dependencies and inter-class

* Corresponding author.
E-mail addresses: liuxinda@nwu.edu.cn (X. Liu), zhangqinyu@stu.nwu.edu.cn (Q. Zhang), minweiqing@ict.ac.cn (W. Min), ghgeng@nwu.edu.cn (G. Geng), sqjiang@ict.ac.cn (S. Jiang).

https://doi.org/10.1016/j.patcog.2026.113610
Received 25 August 2025; Received in revised form 4 March 2026; Accepted 22 March 2026
Available online 25 March 2026

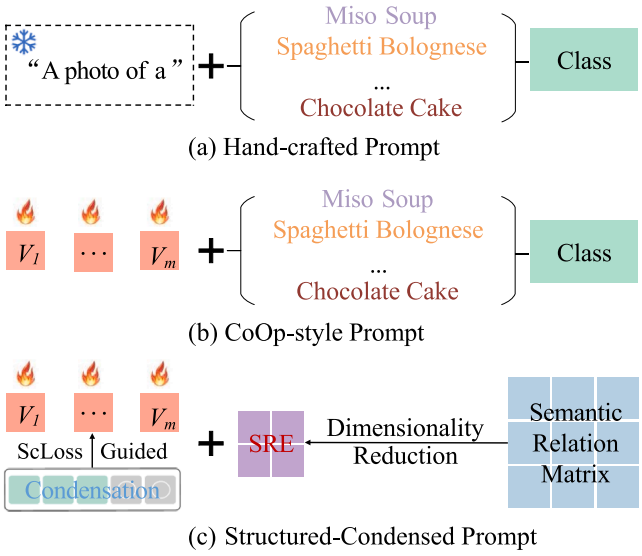

Fig. 1. From static to semantically structured prompts: a paradigm shift in prompting mechanisms.

correlations, which are crucial for distinguishing visually similar categories. The challenge is particularly pronounced in fine-grained classification, where effective recognition hinges on understanding the intricate semantic structure underlying label relationships.

We address this limitation by introducing **Structured-Condensed Prompt Tuning** (SCPT), which formulates a structure-aware semantic alignment mechanism within the CoOp-style prompt tuning framework. Instead of relying on discrete class tokens, SCPT embeds Semantic Relation Encoding (SRE) into the prompt space, preserving the global semantic topology among labels through pairwise relational distances, as shown in Fig. 1(c). This enables vision-language models to reason over structured category representations, where each category embedding explicitly encodes its semantic relations to other categories in a shared embedding space, rather than memorizing disconnected labels. We further propose Semantic Condensation loss (ScLoss) to enhance semantic focus by suppressing intra-class noise and amplifying discriminative cues essential for classification. Together, SRE and ScLoss form a unified framework that captures both inter-class structure and intra-class focus, enhancing both few-shot adaptation and base-to-novel generalization. Experimental results show that SCPT consistently outperforms state-of-the-art prompt tuning methods, achieving substantial gains in both few-shot learning and base-to-novel generalization.

In summary, the key contributions of this paper are as follows:

(1) We propose a structure-aware prompt tuning method, Structured-Condensed Prompt Tuning (SCPT), which enhances the semantic structure modeling in vision-language models for FGIR. This approach explicitly captures the inter-class semantic relationships and improves the model's ability to distinguish subtle differences between visually similar categories.

(2) We present SRE to model the inter-class semantic topology by encoding structured label relationships. This method leverages pairwise relational distances to preserve the global semantic structure, enabling the model to better understand the hierarchical dependencies and correlations between classes.

(3) We design ScLoss to suppress redundant supervision signals and extract discriminative components from the global semantic space. This loss function enhances the model's focus on task-relevant semantics, improving both few-shot adaptation and generalization to novel classes.

(4) We conduct extensive experiments on 14 diverse FGIR benchmarks. SCPT demonstrates superior performance compared to existing prompt tuning methods. It achieves significant gains in both few-shot learning and base-to-novel generalization tasks, establishing a new state-of-the-art in this domain.

2. Related works

2.1. Vision-language models (VLMs)

VLMs such as CLIP [12], ALIGN [15], and BLIP [16] leverage large image-text datasets to achieve superior discriminative power and generalization over traditional single-modality models. By aligning visual and textual information through self-supervised learning, these models excel in zero-shot tasks and transfer their learned representations to downstream applications like image retrieval [17,18], image segmentation [19,20], and visual question answering [21]. Although CLIP has shown strong zero-shot performance, fine-tuning for specific tasks can enhance its effectiveness, especially in resource-limited settings. Recent works [13,14] have explored fine-tuning VLMs for few-shot image recognition, addressing challenges like data scarcity and computational constraints. In this paper, we propose a novel text-prompt tuning method to adapt CLIP for few-shot fine-grained visual recognition, enhancing the model's flexibility and applicability.

2.2. Prompt tuning for VLMs

Prompt tuning seeks to enhance the adaptability of VLMs by introducing a limited number of learnable parameters, thereby allowing the generic features acquired during pre-training to be better aligned with relevant downstream tasks. The fundamental concept is to integrate category information through learnable textual or visual prompts, replacing handcrafted templates such as "a photo of a classname". These prompts are optimized via backpropagation, while the pre-trained model remains frozen. Existing prompt tuning methodologies often improve model performance by leveraging learnable parameters across various modalities while incorporating additional semantic information. For instance, CoOp[13] replaces the handcrafted templates in CLIP with learnable soft prompts that integrate class names, resulting in substantial advancements in few-shot classification tasks. CoCoOp[14] further enhances the generalization capability of the CoOp model by merging image features with textual soft prompts. Moreover, methods such as MaPle [22], and PromptSRC[23] also introduce learnable visual prompts to bolster model performance. More recently, DGPrompt [24] proposes a dual-guidance prompt generation framework that jointly exploits textual semantic cues and visual structural information to enhance cross-modal alignment in vision-language models. ProDa [25] optimizes the direction of prompt updates while discarding conflicting adjustments to constrain the semantic differences between learnable text features and general text semantics. Conversely, KgCoOp [26] limits the distance between learnable text features and their general semantic counterparts, ensuring that specific semantics remain closely aligned with broader concepts. Building upon KgCoOp, TCP [27] integrates class-related general semantic knowledge extracted from CLIP into deeper layers of the text encoder, thereby enhancing the encoder's capacity to differentiate between classes. Related to structured semantics, graph-based methods rely on explicit graph construction and additional supervision to model inter-class relationships, whereas SCPT leverages the implicit semantic structure of CLIP's pretrained embedding space without introducing extra graphs or annotations.

Despite the advancements made by existing prompt tuning approaches in fine-grained image recognition, many still utilize original categories as classification labels, inadequately addressing the semantic limitations inherent in FGIR tasks. Beyond prompt-based adaptation, recent work has explored test-time adaptation for vision-language models. Bayesian Test-Time Adaptation [28] frames adaptation as Bayesian inference to handle distribution shifts at inference time, but does not explicitly model the semantic granularity or structural relationships required for fine-grained recognition. Therefore, when processing

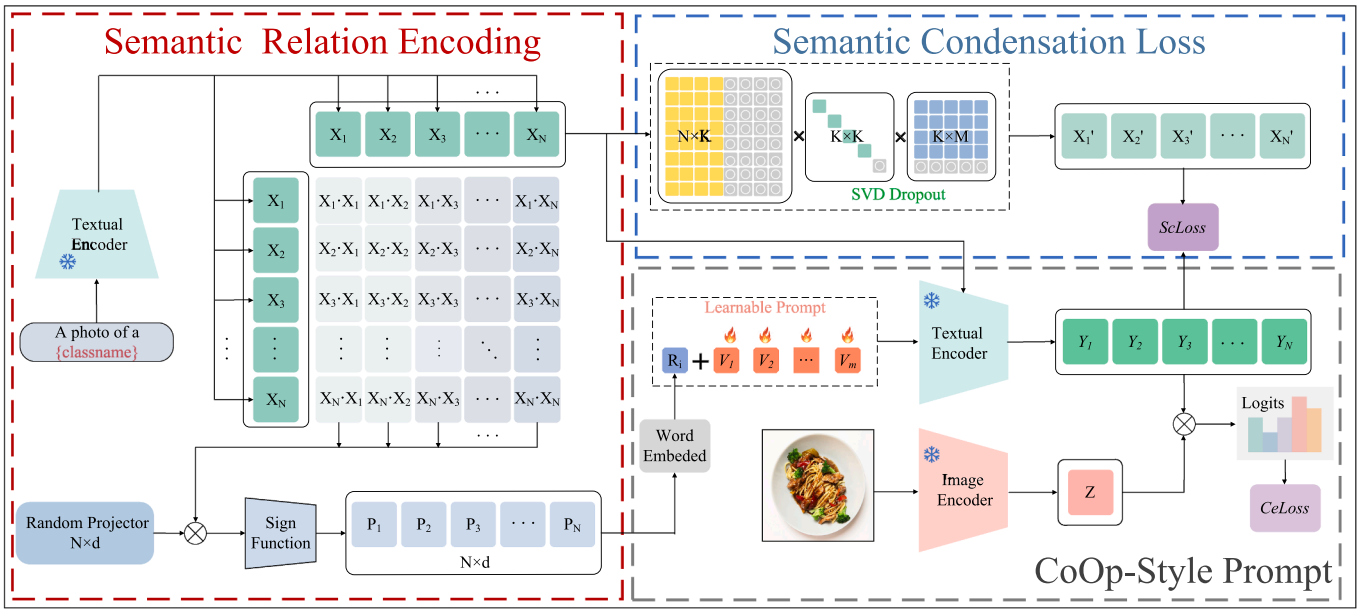

Fig. 2. Framework of the proposed Structured-Condensed Prompt Tuning (SCPT) method. Given a class name, learnable prompts are encoded via a textual encoder. The Structural Relation Encoder (SRE) constructs a similarity matrix through random projection and pairwise feature interactions. The Structural Contrastive Loss (ScLoss) performs low-rank approximation and SVD dropout to suppress task-irrelevant noise. Final classification logits are obtained by aligning visual features and textual features via contrastive learning. The entire framework is jointly optimized using Cross-Entropy Loss (CeLoss) and Structural Contrastive Loss (ScLoss).

fine-grained categories, CLIP may lack sufficient associative semantics learned from its training data, leading to suboptimal generalization performance. In contrast, the proposed method explicitly models semantic structural relationships and enhances FGIR performance by extracting discriminative components from the global semantic space.

3. Method

The proposed method is grounded in the observation that employing isolated, discrete category labels within prompt templates for prompt tuning underutilizes the rich semantic information inherent in these labels. This drawback is especially pronounced in fine-grained image recognition tasks. To mitigate this issue, we introduce Structured-Condensed Prompt Tuning (SCPT), which enhances the semantic expressiveness of fine-grained label prompts by incorporating Semantic Relation Encoding (SRE) alongside a Semantic Condensation loss (ScLoss), as illustrated in Fig. 2. In the following sections, we first provide a brief review of the CoOp-style prompt method, followed by a detailed description of the proposed SCPT framework.

3.1. CoOp-style prompt tuning

Existing CoOp-style prompt tuning algorithms utilize the powerful CLIP as their backbone. CLIP consists of a Transformer-based text encoder and a vision encoder based on either Vision Transformers [29] or Residual Networks [30], where the vision encoder generates visual embeddings from images, and the text encoder transforms text prompts into textual embeddings. During training, CLIP employs contrastive loss to align these embeddings, facilitating effective zero-shot inference. For zero-shot classification, CLIP compares image features with class weights generated from descriptive text, enabling classification without requiring specific category training. Formally, let the image feature extracted by the image encoder be denoted as $\mathbf{Z}$, and let $\mathbf{X} = \{X_i\}_{i=1}^{N}$ represent the class weight vectors generated by the text encoder, where N is the number of classes. Each class is associated with a handcrafted prompt template, which is first transformed into a vectorized textual token using the word embedding function $e(.)$. Specifically, for the ith class, the textual token is given by:

$$t_i = e(\text{“a photo of [CLASSNAME]”}). \tag{1}$$

Subsequently, the text encoder θ maps these vectorized textual tokens into class-level embeddings:

$$X_i = \theta(t_i). \tag{2}$$

Finally, the prediction probability is defined as:

$$p(y=i|x) = \frac{\exp\left(\frac{\text{sim}(X_i, Z)}{\tau}\right)}{\sum_{j=1}^{N} \exp\left(\frac{\text{sim}(X_j, Z)}{\tau}\right)}, \tag{3}$$

where τ is a temperature parameter learned by CLIP and $\text{sim}(\cdot,\cdot)$ denotes cosine similarity.

To enhance CLIP's performance on downstream tasks, CoOp replaces the handcrafted prompt template with learnable prompts. By training on a small number of label-relevant images, CoOp improves the discriminability of class-specific textual embeddings. Specifically, CoOp introduces m learnable context vectors $V = \{v_1, v_2, \ldots, v_m\}$. The class label embedding c_i for the ith class is concatenated with the context vectors to form the prompt $V_i^* = \{c_i, v_1, v_2, \ldots, v_m\}$. This prompt V_i^* is then fed into the text encoder θ, yielding the textual class embedding $Y_i = \theta(V_i^*)$. Finally, the textual embeddings for all classes are defined as $\mathbf{Y} = \{Y_i^{coop}\}_{i=1}^{N}$.

CoOp optimizes the learnable context vectors V by minimizing the contrastive loss between the image embedding z and its corresponding class embedding $\mathbf{Y}$:

$$L_{ce} = -\sum_{\mathbf{z}\in\mathbf{Z}} \log \frac{\exp\left(\text{sim}\left(\mathbf{Z}, \mathbf{Y}_{gt}\right)/\tau\right)}{\sum_{i=1}^{t} \exp\left(\text{sim}\left(\mathbf{Z}, \mathbf{Y}_i\right)/\tau\right)}, \tag{4}$$

where gt is the corresponding label of the image embedding, $\mathbf{Z}$ is the complete set of embeddings.

To further enhance textual semantics and improve the discriminability of CoOp, previous works [27] have explored refining textual embeddings by reducing their dimensionality and incorporating class-aware tokens into the text encoder. Building on these approaches, SCPT applies this enhancement within the text encoder, facilitating the generation of more discriminative textual representations.

3.2. Structured-condensed prompt tuning

Unlike existing CoOp-style prompt tuning methods, the proposed SCPT framework enhances the modeling of label semantic structure by jointly leveraging two complementary components. SRE explicitly captures the inter-class semantic topology, replacing isolated category labels with structure-aware embeddings that encode pairwise semantic relations among categories. ScLoss compresses supervision signals by suppressing redundancy and extracting discriminative components from the global semantic space. As illustrated in Fig. 2, these two modules work in tandem to improve semantic alignment and fine-grained discriminability. We detail each component in the following sections.

3.2.1. Semantic relation encoding

In this work, we define structured category representations as category embeddings that preserve global inter-class semantic topology, rather than treating each category as an independent token. The CLIP model, with its strong zero-shot generalization capabilities, is employed to extract semantic features for category classification. For fine-grained image recognition tasks with N categories, we construct category-specific prompts c_i in the form "a photo of a c_i", which are passed through CLIP's text encoder to generate category embeddings, denoted as $\mathbf{X}$. Each X_i represents the semantic feature of the ith category. To quantify semantic relation between categories, we compute a similarity matrix using cosine similarity:

$$S_{ij} = \frac{\mathbf{X}_i \cdot \mathbf{X}_j}{\|\mathbf{X}_i\|\|\mathbf{X}_j\|}, \quad \forall i, j \in \{1, 2, \dots, N\}, \tag{5}$$

where $\mathbf{X}_i$ and $\mathbf{X}_j$ are the feature vectors of the ith and jth categories. Each row S_i in the similarity matrix captures the semantic similarity between the ith category and all others. This design stems from the research that CLIP's text embeddings inherently encode taxonomic relationships, and preserving such relations through similarity constraints Eq. (5) helps prevent semantic distortion during prompt tuning.

However, due to CLIP's text prompt length limitations [31], directly using the similarity matrix S as category labels is not feasible. Inspired by the Johnson-Lindenstrauss (JL) Lemma [32], which states that a set of N points in a high-dimensional space can be embedded into a lower-dimensional space $\mathbb{R}^d$ ($d \geq \mathcal{O}(\epsilon^{-2} \log N)$) while approximately preserving pairwise distances, we propose a signed random projection method for compact semantic representation. Given a similarity matrix $S \in \mathbb{R}^{N \times N}$, we generate a binary embedding $P = \{P_i\}_{i=1}^{N} \in \{0, 1\}^{N \times d}$ as follows:

$$P = \text{sign}\left(SW^{\top}\right), \tag{6}$$

where $W \in \mathbb{R}^{d \times N}$ is a random projection matrix whose entries are drawn i.i.d. from $\mathcal{N}(0, 1)$. Each row of W defines a hyperplane that randomly partitions the semantic space, and the sign operation encodes whether category embeddings lie on the positive side of these hyperplanes. This binarization compresses semantic relations into a Hamming space where similar categories share overlapping bit patterns. The sign function sign$(\cdot)$ applies element-wise binarization, mapping positive values to 1 and non-positive values to 0. Geometrically, each row vector $W_i \in \mathbb{R}^{1 \times N}$ defines a randomly oriented hyperplane that partitions the semantic space, and $P_i \in \{0, 1\}^d$ encodes the relative positioning of categories with respect to W_i. According to JL Lemma, the number of parameter d is defined as $d = \lceil \log_2(N) + d_{free} \rceil$, where d_{free} introduces additional space to enhance randomness. This ensures the vectors remain independent and non-repetitive after dimensionality reduction.

Similar to the CoOp approach, we first input the binary encoding P into the word embedding function $e(\cdot)$ to generate the SRE, denoted as $R = \{R_i^N\}$, where $R_i = e(P_i)$. Subsequently, we combine the SRE with a learnable context vector to construct a semantic distance-aware prompt vector:

$$V_i = \{R_i, v_1, v_2, \dots, v_m\}. \tag{7}$$

This approach efficiently encodes relational semantics while preserving inter-class structure in the lower-dimensional space.

3.2.2. Semantic condensation loss

KgCoOp [26] demonstrates that aligning learnable prompts with handcrafted prompt-generated embeddings via an MSE regularizer helps extract discriminative components from the global semantic space, thereby improving classification performance. However, this approach may simultaneously introduce redundant or noisy signals–particularly problematic in fine-grained settings, where class semantics are narrowly scoped and visually diverse. Such coarse supervision can mislead the optimization process and hinder the learning of compact, class-specific representations. Motivated by this hypothesis, we propose a straightforward yet effective strategy: Specifically, let $\mathbf{X} \in R^{N \times M}$ denote the text embeddings generated from handcrafted prompts and $\mathbf{Y} \in R^{N \times M}$ represent the embeddings obtained from learnable prompts. Treating $\mathbf{X}$ as a semantic data matrix and applying singular value decomposition (SVD) to remove noise. By preserving dominant singular values and eliminating smaller ones, we aim to denoise the embeddings. Specifically, we perform SVD on $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}. \tag{8}$$

Here, $\mathbf{\Sigma} = (\sigma_1, \sigma_2, \dots, \sigma_r)$ represents the singular values arranged in descending order, i.e., $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq 0$, where r denotes the rank of the matrix. We retain only the top K singular values, yielding:

$$\begin{cases} \mathbf{U}_K = \mathbf{U}_{[:,:K]} \in \mathbb{R}^{N \times K}, \\ \mathbf{\Sigma}_K = \text{diag}(\mathbf{\Sigma}_{[:K]}) \in \mathbb{R}^{K \times K}, \\ \mathbf{V}_K^{\top} = \mathbf{V}_{[:K,:]}^{\top} \in \mathbb{R}^{K \times M}. \end{cases} \tag{9}$$

The denoised embedding matrix $\mathbf{X}'$ is then reconstructed as:

$$\mathbf{X}' = \mathbf{U}_K \mathbf{\Sigma}_K \mathbf{V}_K^{\top}. \tag{10}$$

Our goal is to mitigate the model's tendency to forget general semantics by minimizing the divergence between inter-class and generic semantic knowledge:

$$L_{\text{sc}} = \frac{1}{N} \sum_{i=1}^{N} \|Y_i - X_i'\|_2^2, \tag{11}$$

where $\| \cdot \|_2$ is the Euclidean norm, and N_c is the number of seen classes. Finally, we integrate the standard contrastive loss L_{ce} with our Semantic Condensation loss L_{sc} to define the overall objective function:

$$L = L_{\text{ce}} + \lambda L_{\text{sc}}, \tag{12}$$

where λ controls the contribution of L_{sc} within the total loss.

As illustrated in Fig. 3(a), applying this denoising process improves the MSE performance compared to using the original embeddings. However, the results also indicate that the choice of K significantly influences the outcome across different datasets. This suggests that SVD-based denoising is highly sensitive to the selection of K, raising an important question: **Can a robust mechanism be devised to determine the optimal value of K?**

3.2.3. Optimal value of K

To explore this question, we visualize the distribution of nonzero singular values after performing SVD on semantic embeddings from various fine-grained datasets, as shown in Fig. 3. Compared to Fig. 3 (a), a consistent pattern emerges: despite differences in matrix rank, the optimal number of retained singular values tends to fall between the mid-range and lower-range singular values. This suggests that smaller singular values predominantly capture noise, with the smallest ones being particularly susceptible to noise-induced distortions.

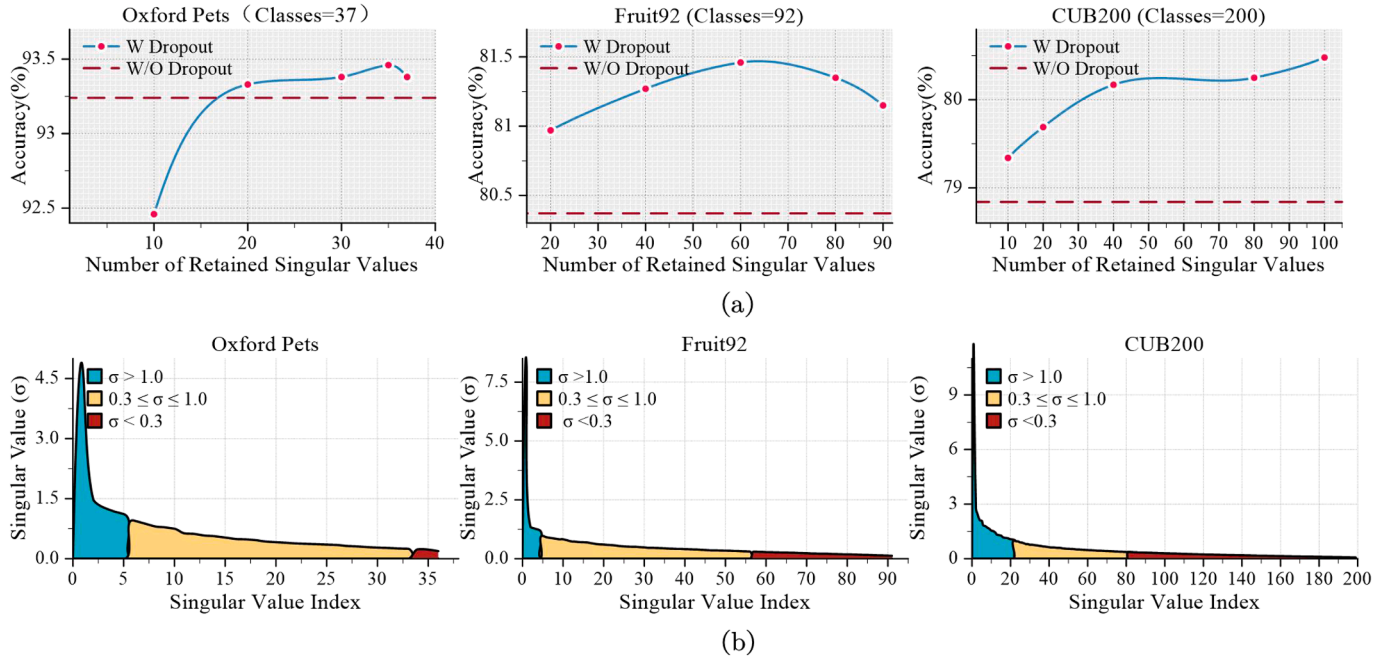

Fig. 3. Comparative Analysis of SVD Dropout Efficacy: Accuracy Dynamics and Singular Value Distribution. (a)Accuracy variation across three datasets under varying principal component retention levels K. Dashed red lines denote baseline performance without SVD Dropout. (b)Post-decomposition singular value magnitude distribution of semantic embeddings. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

To further substantiate this observation, We assume that $\mathbf{X} \in \mathbb{R}^{N\times M}$ is a purely noisy Gaussian random matrix, where each element is independently and identically distributed (i.i.d.) with zero mean and variance σ^2. The covariance matrix of $\mathbf{X}$ can be expressed as:

$$\mathbf{S} = \frac{1}{M} X X^T. \tag{13}$$

According to the Marchenko-Pastur theorem [33], when $N, M \to \infty$, and $\frac{M}{N} \to r \in (0, 1]$. The singular values of the sample covariance matrix follow a characteristic distribution given by:

$$p(\sigma) = \frac{1}{2\pi r\sigma^2}(\sigma_{\max} - \sigma)(\sigma - \sigma_{\min}), \quad \sigma_{\min} \le \sigma \le \sigma_{\max}, \tag{14}$$

where $\sigma_{\min}$ and $\sigma_{\max}$ denote the lower and upper bounds of the singular value spectrum, determined by the data dimensions N, M and the noise variance σ^2. Notably, the density of singular values near $\sigma_{\min}$ is markedly high, indicating that most singular values in a pure noise matrix reside in this region. Conversely, singular values exceeding $\sigma_{\max}$ are statistically improbable under a pure noise assumption, implying that they are more likely to correspond to meaningful semantic structures rather than random fluctuations.

Building on this theoretical analysis, we propose an adaptive filtering mechanism that removes singular values below a predefined threshold, effectively reducing noise while preserving as much semantic information as possible. Notably, the optimal value of K tends to shift slightly to the right of the theoretical optimum. This shift occurs because some smaller singular values may still carry semantic information. To address this, we design a probabilistic selection function $P(\sigma_i)$ as follows:

$$P(\sigma_i) = \begin{cases} 1, & \text{if } \sigma_i \ge \tau, \\ e^{(\sigma_i - \tau)^2 - N_{\text{below}}}, & \text{otherwise.} \end{cases}, \tag{15}$$

where τ represents the threshold, and N_{below} denotes the number of singular values below τ. This formulation ensures that singular values exceeding the threshold are always retained, while those closer to the threshold have a higher probability of being preserved. The term $e^{-N_{\text{below}}}$ serves as a regularization factor, reflecting the observation that a larger number of small singular values increases the likelihood of noise. Consequently, when more small singular values are present, the probability of selecting any individual one is reduced, reinforcing the suppression of noise-dominated components. Thus, the final selection of K is determined as the index of the i-th singular value that is retained based on the probabilistic selection function, formulated as:

$$K = \arg\max_i \{P(\sigma_i) \ne 0\}. \tag{16}$$

This ensures that K corresponds to the largest index among the selected singular values, striking a balance between noise reduction and semantic preservation.

Algorithm 1: Training of structured-condensed prompt tuning (SCPT).

Input : Training data $\{(x, y)\}$, class names $\{c_i\}_{i=1}^N$, frozen encoders $f_{\text{img}}, f_{\text{txt}}$, projection matrix W, hyperparameters λ, τ
Output: Optimized context vectors V
// Initialization
Encode handcrafted prompts to obtain text features X;
Compute class similarity matrix S from X;
Generate Semantic Relation Encoding (SRE):
$P = \text{sign}(SW^\top)$, $R_i = e(P_i)$;
// Optimization
while *until convergence* **do**
  Construct structured prompts $V_i = \{R_i, V\}$ and encode $Y_i = f_{\text{txt}}(V_i)$;
  Extract image features $z = f_{\text{img}}(x)$;
  Compute logits via cosine similarity;
  Compute contrastive loss $\mathcal{L}_{\text{ce}}$;
  Compute Semantic Condensation loss $\mathcal{L}_{\text{sc}}$ via SVD-based filtering with threshold τ;
  Update V by minimizing $\mathcal{L}_{\text{ce}} + \lambda\mathcal{L}_{\text{sc}}$;

SCPT introduces a structured prompt tuning strategy that jointly leverages inter-class relational encoding and intra-class semantic compression. SRE encodes inter-class relations via randomized hyperplane

Table 1
Comparing SCPT to other prompt tuning methods in few-shot learning (%).

Method	Dog Breed	Oxford Pets	Oxford Flowers	Stanford Cars	Web Cars	Stanford Dogs	Fruit92	CUB200	Food101	Food172	Food200	Food500	FGVC Aircraft	Veg200	Avg.
CLIP	62.30	89.10	70.70	65.70	63.80	64.50	41.90	54.70	85.90	30.20	49.60	42.50	24.90	35.90	55.84
CoOp	75.45	93.16	95.83	77.72	70.52	73.02	72.39	70.91	86.40	60.60	57.78	47.40	38.69	61.21	70.08
CoCoOp	72.06	93.89	90.88	71.58	69.07	70.46	63.70	64.04	87.38	44.60	54.19	45.00	34.22	50.36	65.11
KgCoOp	71.87	92.96	92.08	72.14	70.21	67.56	67.21	65.79	87.11	50.96	56.00	45.98	34.10	53.53	66.25
TCP	79.79	92.99	97.45	83.70	73.00	74.45	80.43	78.58	87.32	73.63	60.82	54.90	44.56	77.2	75.63
ProText	61.86	92.72	72.42	66.77	61.61	55.04	47.76	54.32	85.35	35.51	46.49	35.60	29.01	34.38	55.63
ATPrompt	76.44	93.32	96.97	79.05	70.71	73.89	76.75	74.55	85.73	65.73	57.89	48.90	40.37	67.40	71.98
SCPT	80.81	93.38	97.72	86.21	74.48	75.58	81.27	80.48	87.45	75.34	60.88	55.00	46.41	78.72	76.70

projections that preserve CLIP's semantic topology, while ScLoss distills task-critical semantics through singular value truncation. Algorithm 1 provides a unified view of how SRE and ScLoss are integrated into the SCPT framework. Their coordinated operation, with SRE expanding inter-class discriminability and ScLoss contracting intra-class redundancy, establishes a robust foundation for fine-grained adaptation.

4. Experiments

To evaluate the effectiveness of the proposed method, we follow standard practice on a series of downstream tasks, including few-shot experiments, base-to-novel generalization, and ablation study.

4.1. Experimental setup

Datasets: We employed 14 fine-grained classification datasets including Dog Breed [34], Oxford Pets [35], Oxford Flowers [36], Stanford Cars [37], Web Cars [38], Stanford Dogs [39,40] , Fruit92 [41] , CUB200 [42], Food101 [43], Food172 [44], Food200 [45], Food500[46], FGVC Aircraft [47] and Veg200 [41].
Implementation Details: For all experiments, we utilize ViT-B/16 as the vision backbone. In the SCPT model, the context length for learnable prompts is fixed at 4, and the context vectors are initialized from a Gaussian distribution ($\sigma = 0.02$), threshold τ is 0.3. The hyperparameter λ is 8.0, same as KgCoOp. The models are trained on the RTX 4090 GPU with 24GB memory using the SGD optimizer. The training process spans 50 epochs with a batch size of 32. The initial learning rate is set to 2e-3 and dynamically adjusted according to the cosine annealing schedule. All comparisons are conducted under a consistent experimental protocol, with method-specific settings applied only when required by the original implementation.
SOTA Methods: We compared SCPT with the zero-shot inference capability of CLIP [12] (ICML 2021) and conducted a comprehensive comparison with recent textual prompt tuning methods. e.g., CoOp [13] (IJCV 2022), CoCoOp [14] (CVPR 2022), KgCoOp [26] (CVPR 2023), TCP [27] (CVPR 2024), ProText [48] (AAAI 2025) and ATPrompt [49] (ICCV 2025).

4.2. Few-shot learning evaluations

To validate that the SCPT method offers enhanced fine-grained semantic distinction compared to existing CoOp-style prompt templates, we conducted few-shot classification across all 14 fine-grained image datasets using K-shot labeled source images. Evaluations were performed on a standard test domain within the same class space as the training set.

Table 1 reports the performance of SCPT on 14 fine-grained datasets under the 16-shot setting, achieving an average accuracy of 76.70%. Compared with the state-of-the-art method TCP, SCPT consistently outperforms it on all benchmarks except Oxford Pets. This dataset contains the fewest fine-grained categories, which constrains the expressiveness of the induced semantic topology and thus limits the potential gains from semantic-aware tuning. Overall, SCPT achieves an average improvement of more than 1% across all datasets.

4.3. Base-to-novel generalization

to ensure consistency with prior studies such as TCP, we partition each dataset into two disjoint sets: base classes and novel classes, following a zero-shot learning framework where base and novel classes do not overlap. This setup enables us to quantify the model's generalization capability to fine-grained, out-of-domain categories. For each evaluation stage, SRE is computed from the textual semantic representations of the candidate class names and projected using the same random projection function as in training; the resulting SRE is then combined with the class name and the learnable context vectors to form the final textual prompt.

Table 2 compares SCPT with CLIP and recent prompt tuning methods under the 16-shot setting. SCPT outperformed existing approaches, showing improvements on both base and novel classes. The SRE-enhanced prompt captured semantic relation, yielding a 1.10% average gain over TCP across 14 datasets. The incorporation of Semantic Condensation loss (ScLoss) mitigated irrelevant semantic interference, resulting in a 71.15% harmonic mean, demonstrating strong generalization to novel classes. In contrast to CoOp methods, which often overfit to base classes, SCPT achieved an 77.84% improvement in harmonic mean performance, a 9.79% increase in novel class generalization, and a 2.46% boost in base class accuracy. These results highlight SCPT's superior ability to balance generalization and base class performance, achieving the highest overall accuracy across all 14 datasets.

4.4. Ablation study

Effects of SRE and ScLoss. The proposed SCPT method integrates two core components: SRE and ScLoss. SRE encodes class names to reflect inter-class relationships and enhance class distinction, while ScLoss suppresses irrelevant semantic information to improve robustness. To ensure a fair and controlled evaluation of the proposed components, we adopt CoOp-style prompts with semantic regularization following the TCP formulation as the baseline setting in the ablation study. This design choice allows us to isolate the individual contributions of SRE and ScLoss under a consistent semantic regularization framework, avoiding performance variations caused by changes in the underlying prompt formulation.

As shown in Table 3, SRE and ScLoss contribute to performance improvement through different mechanisms. SRE enhances class separability by encoding fine-grained inter-class similarity structures, but its reliance on semantic relations makes it more vulnerable to semantic noise. In contrast, ScLoss improves robustness by suppressing irrelevant variations via low-rank filtering, yet lacks the capacity to explicitly capture structural relations among classes. When combined, the two components complement each other–SRE promotes discriminability, while

Table 2
Comparison with SOTA methods on base-to-novel generalization (%) in 16-shots. HM: Harmonic mean.

(a) Average results

Method	Base	Novel	HM
CLIP	60.09	63.88	61.93
CoOp	76.26	55.12	63.99
CoCoOp	71.54	62.86	66.92
KgCoOp	72.39	63.50	67.65
TCP	77.62	64.54	70.48
ProText	60.34	61.91	61.12
ATPrompt	76.85	56.10	64.86
SCPT	78.72	64.91	71.15

(b) Dog Breed

Method	Base	Novel	HM
CLIP	71.80	71.60	71.70
CoOp	87.71	58.89	70.47
CoCoOp	83.75	74.39	78.79
KgCoOp	79.32	63.09	70.28
TCP	**87.73**	74.88	80.80
ProText	72.55	71.00	71.77
ATPrompt	87.08	67.16	75.83
SCPT	87.66	**75.30**	**81.01**

(c) Oxford Flowers

Method	Base	Novel	HM
CLIP	72.08	**77.80**	74.83
CoOp	97.63	69.55	81.23
CoCoOp	94.87	71.75	81.71
KgCoOp	95.00	74.73	83.65
TCP	97.73	75.57	85.23
ProText	74.36	78.44	76.35
ATPrompt	98.02	69.28	81.18
SCPT	**98.08**	75.89	**85.57**

(d) Oxford Pets

Method	Base	Novel	HM
CLIP	91.17	97.26	94.12
CoOp	94.24	96.66	95.43
CoCoOp	**95.20**	97.69	**96.43**
KgCoOp	94.65	97.76	96.18
TCP	94.67	97.20	95.92
ProText	94.95	98.00	96.45
ATPrompt	95.61	97.04	96.32
SCPT	94.76	**97.87**	96.29

(e) Stanford Cars

Method	Base	Novel	HM
CLIP	63.37	74.89	68.65
CoOp	76.20	69.14	72.49
CoCoOp	70.49	73.59	72.01
KgCoOp	71.76	75.04	73.36
TCP	80.80	74.13	77.32
ProText	64.54	76.08	69.84
ATPrompt	77.27	66.54	71.50
SCPT	**81.17**	**75.66**	**78.32**

(f) Fruit92

Method	Base	Novel	HM
CLIP	53.00	61.10	56.76
CoOp	85.39	46.96	60.60
CoCoOp	74.58	60.05	66.53
KgCoOp	76.73	57.37	65.65
TCP	85.83	63.79	73.19
ProText	54.63	52.70	53.65
ATPrompt	85.92	47.21	60.94
SCPT	**86.28**	**64.34**	**73.71**

(g) CUB200

Method	Base	Novel	HM
CLIP	63.90	51.90	57.28
CoOp	81.48	37.37	51.24
CoCoOp	74.46	51.20	60.68
KgCoOp	75.62	52.60	62.04
TCP	82.10	52.97	64.39
ProText	65.07	50.70	56.99
ATPrompt	82.02	40.96	54.64
SCPT	**82.96**	52.67	**64.43**

(h) Food101

Method	Base	Novel	HM
CLIP	90.10	91.22	90.66
CoOp	89.44	87.50	88.46
CoCoOp	**90.70**	91.70	91.09
KgCoOp	90.50	91.70	91.09
TCP	90.57	91.37	90.97
ProText	90.20	91.98	91.08
ATPrompt	88.86	88.95	88.90
SCPT	90.69	91.54	**91.15**

(i) Veg200

Method	Base	Novel	HM
CLIP	44.90	43.80	44.34
CoOp [13]	78.19	30.20	43.57
CoCoOp	67.19	45.79	54.46
KgCoOp	66.85	44.08	53.13
TCP	83.27	49.53	**62.11**
ProText	44.99	39.20	41.90
ATPrompt	79.85	32.53	46.23
SCPT	**84.23**	48.59	61.63

(j) Web Cars

Method	Base	Novel	HM
CLIP	61.80	73.60	62.45
CoOp	68.04	56.98	65.74
CoCoOp	67.87	73.55	70.86
KgCoOp	68.82	74.31	71.46
TCP	68.70	72.50	70.55
ProText	60.89	72.70	66.27
ATPrompt	67.69	62.27	64.87
SCPT	**69.85**	73.16	**71.47**

(k) Stanford Dogs

Method	Base	Novel	HM
CLIP	61.90	63.00	62.45
CoOp	75.80	58.03	65.74
CoCoOp	74.12	67.87	70.86
KgCoOp	71.24	66.12	68.58
TCP	75.45	68.79	71.97
ProText	63.43	66.80	65.07
ATPrompt	76.5	58.31	66.18
SCPT	**75.89**	**69.67**	**72.65**

(l) FGVC Aircraft

Method	Base	Novel	HM
CLIP	27.19	**36.29**	31.09
CoOp	39.24	30.49	34.30
CoCoOp	33.41	23.71	27.74
KgCoOp	36.21	33.55	34.83
TCP	41.97	34.43	37.83
ProText	30.91	34.13	32.44
ATPrompt	39.80	30.74	34.69
SCPT	**42.57**	35.53	**38.73**

(m) Food200

Method	Base	Novel	HM
CLIP	55.20	59.00	57.03
CoOp	64.96	52.31	57.95
CoCoOp	62.17	**59.87**	61.00
KgCoOp	64.22	59.17	61.59
TCP	66.93	58.93	62.68
ProText	52.25	55.80	53.97
ATPrompt	64.82	52.40	57.95
SCPT	**67.04**	59.22	**62.89**

(n) Food500

Method	Base	Novel	HM
CLIP	48.80	**52.00**	50.35
CoOp	56.95	44.99	50.27
CoCoOp	53.12	51.58	52.34
KgCoOp	55.04	50.25	52.54
TCP	61.37	49.62	54.87
ProText	44.30	43.20	43.74
ATPrompt	57.85	43.91	49.93
SCPT	**61.72**	49.48	**54.93**

(o) Food172

Method	Base	Novel	HM
CLIP	36.00	**40.80**	38.25
CoOp	72.28	32.63	44.96
CoCoOp	49.76	40.04	44.37
KgCoOp	63.22	39.06	48.29
TCP	78.34	39.87	52.85
ProText	31.72	36.00	33.72
ATPrompt	74.62	28.10	40.83
SCPT	**79.06**	39.84	**52.98**

Table 3
Effects of SRE and ScLoss combinations (%). SR:Semantic Relation Encoding. Sc:Semantic Condensation loss. TCP (Reg.) denotes the ablation reference baseline, i.e., CoOp-style prompt tuning with semantic regularization under the same experimental configuration.

Method	*Dog Breed*	*Oxford Pets*	*Oxford Flowers*	*Stanford Cars*	*Web Cars*	*Stanford Dogs*	*Fruit92*	*CUB200*	*Food101*	*Food172*	*Food200*	*Food500*	*FGVC Aircraft*	*Veg200*	*Avg.*
TCP (Reg.)	79.79	92.99	97.45	83.70	73.00	74.45	80.43	78.58	87.32	73.63	60.82	54.90	44.56	77.20	75.63
SR	80.08	93.24	97.39	83.64	72.86	74.87	80.37	78.84	87.52	75.12	60.80	55.09	43.99	77.38	75.80
Sc	79.91	93.33	97.64	84.04	73.30	74.80	80.50	79.65	87.48	74.42	60.92	54.92	44.82	78.08	75.98
SR & *Sc*	80.81	93.38	97.72	86.21	74.48	75.58	81.27	80.48	87.45	75.34	60.88	55.00	46.41	78.72	76.70

ScLoss mitigates noise sensitivity–resulting in an average performance gain exceeding 1%. These results highlight their strong complementarity and validate the effectiveness of joint optimization.

Effect of d_{free} in SRE. This study examines the impact of the additional free space d_{free} on encoding length and recognition accuracy in a 16-shot setting across 14 datasets, as shown in Fig. 4. We evaluate d_{free} values from 0 to 8, with results indicating that $d_{\text{free}} = 3$ yields the best performance. However, the accuracy difference between the highest and lowest values is only 0.36%, suggesting that recognition accuracy is not highly sensitive to

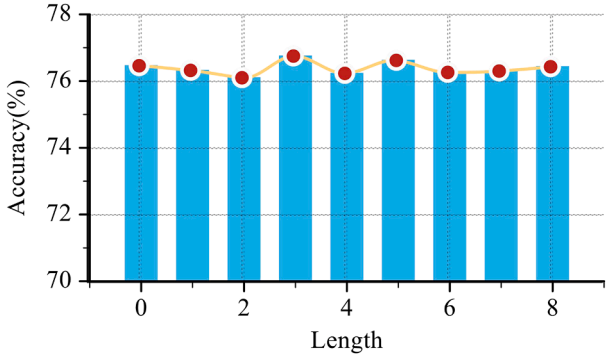

Fig. 4. Impact of additional free space d_{free}.

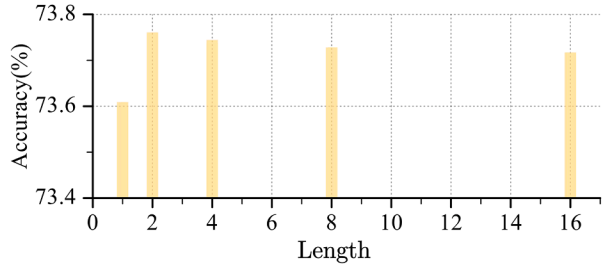

Fig. 5. Effect of learnable prompt Length M.

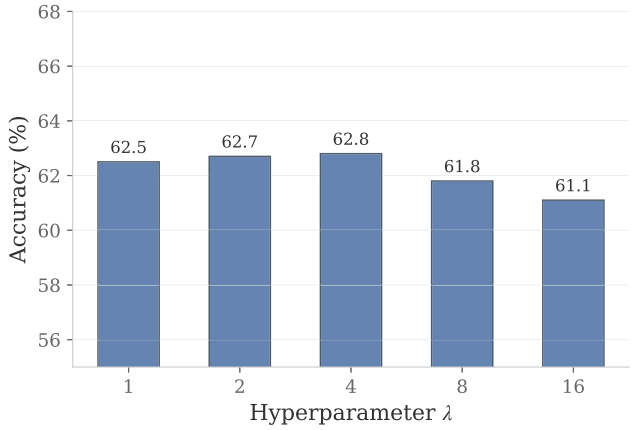

Fig. 6. Effect of loss weight λ.

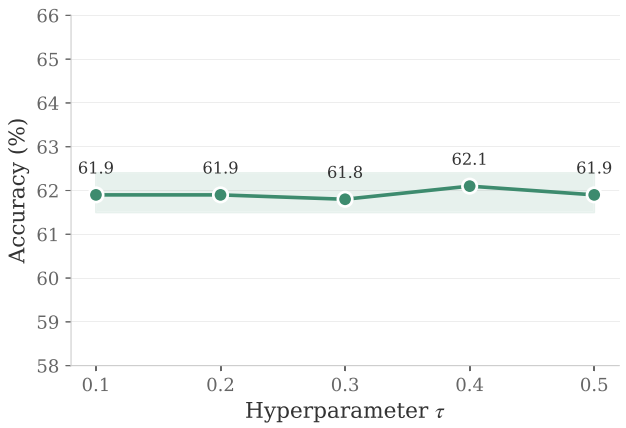

Fig. 7. Effect of threshold τ.

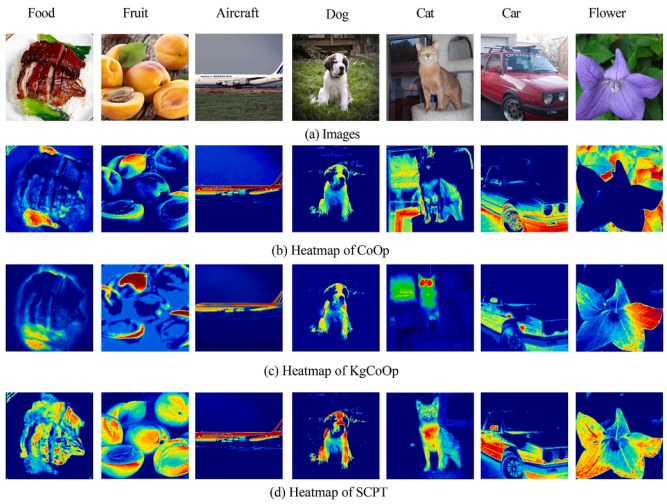

Fig. 8. Comparison of visual attention heatmaps generated by different methods: (a) Ground-truth images; (b) Heatmaps from CoOp similarity scores; (c) Heatmaps from KgCoOp similarity scores; (d) Heatmaps from SCPT similarity scores.

Table 4
Sensitivity analysis of the random projection matrix **W** initialization on the Fruit92 dataset with a ViT-B/32 backbone.

Random Seed	1	2	3	42	100	Mean	Std
Accuracy (%)	75.4	75.3	75.4	75.5	74.9	75.3	±0.23

variations in d_{free}. This implies that the dimensionality reduction is not strongly dependent on the reduced space size in the absence of conflicts.

Effect of Prompt length M. In a 4-shot experiment, we investigate the impact of prompt length M on few-shot performance, as shown in Fig. 5. We evaluate prompt lengths of 1, 2, 4, 8, and 16, finding that SRE-encoded prompts are largely insensitive to learnable parameter variations. While a prompt length of 2 yields the highest accuracy, the difference from the lowest result is just 0.16%. Shorter prompt lengths generally outperform longer ones, indicating that SRE encoding inherently captures more class-specific information, which is particularly advantageous in fine-grained recognition tasks.

Effect of Hyperparameter λ. We evaluate the sensitivity of the loss weight λ by measuring average accuracy on four fine-grained datasets, namely Fruit92, FGVC-Aircraft, CUB200, and Stanford Dogs, using the ViT-B/32 backbone. The value of λ is varied over $\{1, 2, 4, 8, 16\}$. As shown in Fig. 6, the performance remains stable when λ ranges from 1 to 8, with accuracy variations below 1%. The highest accuracy is achieved at $\lambda = 4$, while $\lambda = 8$ yields a comparable result. For consistency with prior work and to avoid dataset-specific tuning, we use $\lambda = 8$ in all experiments.

Effect of Threshold τ. The threshold τ controls the strength of feature noise filtering. We evaluate the sensitivity to τ by varying its value from 0.1 to 0.5 under the same experimental setting, using the ViT-B/32 backbone on Fruit92, FGVC-Aircraft, CUB200, and Stanford Dogs. As shown in Fig. 7, the average classification accuracy remains nearly unchanged across the entire range, with variations below 0.3%. Based on this observation, we use a fixed value of τ in all experiments.

Sensitivity to Random Projection Initialization. We evaluate the impact of initialization on the projection matrix **W** using five random seeds on the Fruit92 dataset (ViT-B/32). As shown in Table 4, SCPT exhibits remarkable robustness, yielding a mean accuracy of 75.3% ±0.23. With a variation of less than 0.6% across all trials, the results conclusively demonstrate that the proposed signed random projection strategy is insensitive to initialization noise, validating its reliability for consistent deployment.

Table 5
Computational efficiency and model complexity analysis on Food-500.

Method	Model Complexity		Computational Efficiency			
	Params (M)	Ratio (%)	Init. (s)	Train (img/s) ↑	Infer. (ms) ↓	Mem. (MB)
CoOp [13]	0.004	0.004	0.058	174.9	183.0	941.4
CoCoOp [14]	0.103	0.100	0.063	6.2	5135.5	942.8
TCP [27]	0.793	0.780	0.127	170.8	264.1	955.6
SCPT (Ours)	0.790	0.770	0.286	140.1	228.4	947.1

Table 6
Performance comparison with different backbones (ViT-B/32 and RN50).

Method	Fruit92	Veg200	Webcar	Aircraft	Dog Breed	Avg.
Backbone: ViT-B/32						
CLIP	48.95	37.42	59.43	20.85	59.35	45.20
CoOp	70.14	58.49	62.90	29.93	70.23	58.34
CoCoOp	30.47	28.90	48.43	7.00	55.07	33.97
KgCoOp	63.22	51.31	63.84	26.70	67.13	54.44
TCP	69.60	64.13	64.00	28.37	68.67	58.95
ProText	42.60	32.13	57.52	10.11	57.17	39.91
ATPrompt	71.40	61.17	62.80	30.60	71.30	59.45
SCPT	**74.40**	**70.74**	**66.97**	**31.93**	**73.13**	**63.43**
Backbone: RN50						
CLIP	46.32	35.14	54.66	18.72	55.90	42.15
CoOp	67.63	54.76	59.43	28.35	67.01	55.44
CoCoOp	37.80	28.40	49.07	16.27	50.83	36.47
KgCoOp	60.63	47.90	59.93	26.05	63.63	51.63
TCP	67.63	63.88	60.94	27.54	66.16	57.23
ProText	38.45	28.70	52.05	6.00	49.08	34.86
ATPrompt	69.47	58.07	59.07	28.33	67.27	56.44
SCPT	**73.57**	**68.53**	**62.99**	**30.62**	**70.63**	**61.27**

Comparison of Computational Efficiency. Table 5 summarizes the complexity analysis on the large-scale Food-500 dataset. Results indicate that while SCPT incorporates the SRE module, it maintains a lightweight design with 0.79M learnable parameters, representing merely 0.77% of the total model capacity. In terms of computational speed, SCPT achieves a training throughput of 140.1 img/s, which is more than 20× faster than CoCoOp and comparable to both CoOp and TCP. During inference, SCPT achieves 228.4 ms per batch, remaining close to TCP in latency while maintaining a similar memory footprint. These results suggest that SCPT maintains a similar parameter scale and computational efficiency to TCP, while providing consistent performance improvements.

Robustness Across Backbones. While our main experiments are conducted with ViT-B/16, we further evaluate SCPT on two additional visual backbones, ViT-B/32 and ResNet-50, to assess its generality across architectures. Experiments are performed on five representative fine-grained datasets: Fruit92, Veg200, WebCar, FGVC Aircraft, and Dog Breed. As shown in Table 6. SCPT achieves consistent state-of-the-art performance on both ViT-B/32 and ResNet-50, exceeding the strongest prior methods by a clear margin in average accuracy. This confirms that the effectiveness of SCPT generalizes well across both transformer- and CNN-based visual backbones.

4.5. Comparison with visual prompt tuning methods

Although SCPT is designed as a purely textual prompt tuning method, we include a comparison with representative recent multimodal prompt tuning approaches to analyze the relationship between prompt complexity and recognition performance. As shown in Table 7, MaPLe and PromptSRC adopt deep prompt injection across nine transformer layers and leverage both visual and textual modalities, whereas SCPT employs a single-layer shallow prompt and operates exclusively in the textual domain.

Table 7
Comparison of prompt learning methods with respect to prompt type, depth, modality, and average accuracy across 15 datasets. *Depth* denotes the number of transformer layers where learnable prompts are injected.

Method	Prompt Type	Depth	Modalities	Avg Acc (%)
MaPLe	Deep	9	Text + Vision	62.79
PromptSRC	Deep	9	Text + Vision	76.94
SCPT (ours)	Shallow	1	Text-only	**76.70**

Despite this substantially lighter design, SCPT attains an average accuracy of 76.70%, which is comparable to PromptSRC in terms of average accuracy and markedly higher than MaPLe across 14 datasets. These results indicate that strong performance can be achieved with shallow, text-only prompt tuning, suggesting that explicitly modeling semantic structure in textual prompts is a viable alternative to deep or multimodal prompt designs.

4.6. Visualization

To investigate the impact of textual prompts on image feature representations, we compute similarity scores between target images and textual prompts using CoOp, KgCoOp, and the proposed SCPT method. Grad-CAM [50] is then applied to backpropagate these similarity scores through the image encoder, producing gradient-based heatmaps that visualize image regions most sensitive to the target similarity scores, as shown in Fig. 8.

Textual embeddings act as explicit supervisory signals during backpropagation, guiding gradient responses toward the target class and facilitating more discriminative region localization. Compared with CoOp and KgCoOp, SCPT produces heatmaps that exhibit stronger focus on target regions, clearer foreground-background separation, and finer class-specific details, while suppressing irrelevant visual information.

5. Limitations

While our proposed framework demonstrates strong effectiveness in text-based prompt learning, several limitations remain. First, our method focuses exclusively on textual prompts and does not incorporate learnable visual prompts. Recent works such as MaPLe [22] and PromptSRC [23] have shown that jointly optimizing textual and visual prompts can further enhance adaptation and generalization in certain scenarios. Extending SCPT to a multimodal prompt learning framework may provide additional representational flexibility, which we leave for future work.

Second, our current design adopts shallow prompt tuning and injects learnable prompts only at a limited number of layers. Although this design choice helps maintain parameter efficiency and training stability, it does not explore deeper prompt mechanisms that interact with intermediate layers of the backbone, as investigated in VPT [51] and DualPrompt [52]. Incorporating deep prompts into the proposed structured and condensed prompt formulation is a promising direction for future research.

Finally, Although SCPT is formulated as a complete prompt tuning model within the CoOp-style framework, extending it to other prompt learning paradigms is an interesting direction for future work.

6. Conclusion

This work revisits prompt tuning for fine-grained recognition from a structural perspective and argues that treating category prompts as isolated tokens fundamentally limits the ability of vision-language models to discriminate subtle inter-class differences. To address this issue, we propose Structured-Condensed Prompt Tuning (SCPT), which explicitly incorporates global semantic structure into textual prompt learning. By introducing Semantic Relation Encoding (SRE), SCPT captures inter-class semantic topology in a lightweight and parameter-efficient manner, while the Semantic Condensation loss (ScLoss) further refines supervision by suppressing redundant or noisy semantic components in the shared embedding space.

Extensive experiments on 14 fine-grained benchmarks demonstrate that SCPT consistently improves both few-shot adaptation and base-to-novel generalization, indicating that structured semantic modeling is particularly beneficial under limited data and distribution shift. The results suggest that incorporating global semantic relations provides a more effective inductive bias than treating category prompts as independent tokens.

Despite these advantages, SCPT has certain limitations. Its effectiveness depends on the availability of a sufficiently rich category set, as semantic topology becomes less informative when the number of classes is small. In addition, the current framework focuses exclusively on textual prompt learning and does not exploit complementary visual or multimodal prompting mechanisms. Future work will explore these directions and extend SCPT to large-scale and open-vocabulary recognition settings.

CRediT authorship contribution statement

Xinda Liu: Writing – original draft, Conceptualization; **Qinyu Zhang:** Writing – review & editing, Writing – original draft, Conceptualization; **Weiqing Min:** Writing – review & editing, Methodology; **Guohua Geng:** Visualization, Funding acquisition; **Shuqiang Jiang:** Project administration, Funding acquisition.

Data availability

All datasets used are publicly available benchmarks. Code will be released upon acceptance.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Beijing Natural Science Foundation (JQ24021), the National Natural Science Foundation of China (Grants No. 62125207 and No. 62472411), the National Key R&D Program of China (Grant No. 2024YFF0907605), and the General Projects of the Shaanxi Provincial Department of Science and Technology (Grant No. 2025JC-YBQN-801).

Appendix A. T-SNE visualizations of visual embeddings

To investigate the impact of SCPT on feature space distribution, we conduct a comparative analysis with the baseline method CoOp using t-SNE visualizations of visual embeddings extracted from test set samples across three datasets with varying numbers of categories: Fruit92 (92 classes), Oxford Flowers (102 classes), and Stanford Cars (196 classes), as shown in Fig. A.1.

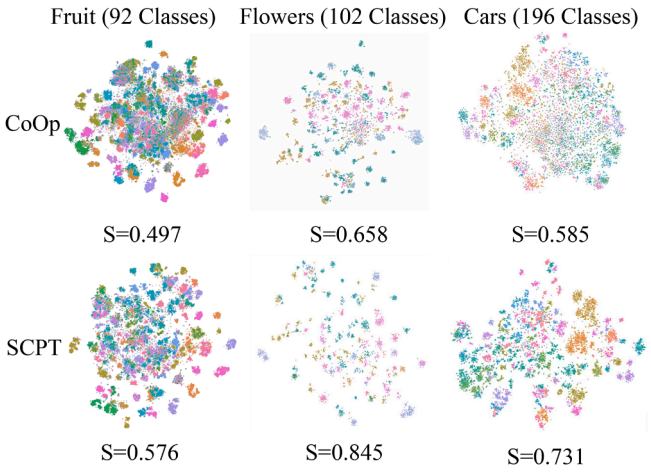

Fig. A.1. t-SNE visualizations of visual embeddings on three datasets. The **Silhouette Coefficient (S)** is reported to quantify cluster compactness and separation.

The Silhouette Coefficient (S) is employed to quantitatively characterize the clustering structure of the visual embeddings, measuring the relative compactness of intra-class samples with respect to inter-class separation. Higher S values indicate more coherent and well-separated clusters.

As illustrated in Fig. A.1, the baseline CoOp exhibits relatively diffuse cluster structures with indistinct boundaries across all evaluated datasets, which is reflected by its lower Silhouette Coefficients. In contrast, SCPT consistently produces more compact intra-class distributions and clearer separation between categories, yielding higher S values. Furthermore, the performance gap between SCPT and CoOp becomes more pronounced as the number of categories increases, indicating that SCPT is effective in maintaining structured feature representations under large-scale and fine-grained recognition settings.

Appendix B. Comparison with conventional few-shot learning settings

Conventional few-shot learning (FSL) methods are typically formulated under the C-way-K-shot episodic paradigm, where models are trained on base classes and evaluated on disjoint novel classes via repeated few-class classification tasks. While widely adopted, this protocol imposes restrictive assumptions on task scale and data availability that diverge from real-world recognition settings.

Specifically, episodic FSL evaluates performance on tasks with a small number of classes (e.g., 5-way or 10-way), whereas practical applications often require discriminating among hundreds of categories simultaneously. Moreover, episodic sampling emphasizes local class subsets and assumes abundant query samples, limiting its ability to reflect global inter-class confusion and large-scale recognition behavior. Most critically, its closed-vocabulary evaluation assesses performance only within individual tasks, offering limited insight into open-vocabulary generalization.

To better align with large-scale and open-world scenarios, we adopt two complementary evaluation protocols commonly used in prompt-based and representation-learning methods: standard few-shot classification and base-to-novel generalization. In the few-shot setting, the model is trained with k labeled samples per class under a global supervised setup and evaluated on a held-out test set, directly measuring scalability under limited supervision. In the base-to-novel setting, training is restricted to base classes, followed by evaluation on both base and novel categories, with generalization quantified by the harmonic mean of their accuracies.

Appendix C. Theoretical analysis of signed random projection

We provide a theoretical justification for the sign-based binarization by showing that the Hamming distance between binary projections consistently approximates the angular distance between the original semantic embeddings. Since vision–language models such as CLIP rely on cosine similarity, this result confirms that our binarization preserves the essential semantic topology of the embedding space.

C.1. Equivalence between hamming distance and angular distance

Lemma 1 (Angular Distance Preservation). *Let $S_i, S_j \in \mathbb{R}^N$ be two non-zero embedding vectors, and let $W \in \mathbb{R}^{d\times N}$ be a random projection matrix whose rows are independently sampled from a spherically symmetric distribution. Define the binary codes $P_i = \text{sign}(WS_i)$ and $P_j = \text{sign}(WS_j)$. The normalized Hamming distance between the two codes is defined as:*

$$d_H(P_i, P_j) = \frac{1}{d}\sum_{k=1}^{d} \mathbb{1}[P_{ik} \neq P_{jk}], \tag{C.1}$$

where $\mathbb{1}[\cdot]$ is the indicator function. This is an unbiased estimator of the normalized angular distance between S_i and S_j, namely:

$$\mathbb{E}[d_H(P_i, P_j)] = \frac{\theta_{ij}}{\pi}, \tag{C.2}$$

where $\theta_{ij} \in [0, \pi]$ denotes the angle between S_i and S_j.

Proof. Consider a single random projection vector w, corresponding to one row of W. The induced hyperplane $\{x : w^\top x = 0\}$ separates S_i and S_j if and only if $\text{sign}(w^\top S_i) \neq \text{sign}(w^\top S_j)$. Due to the rotational invariance of the distribution of w, the probability of this event depends solely on the angle θ_{ij} between the two vectors. It is a classical result in Locality Sensitive Hashing (LSH) that a random hyperplane separates two vectors with probability θ_{ij}/π. Averaging over d independent projections yields $\mathbb{E}[d_H(P_i, P_j)] = \theta_{ij}/\pi$. □

C.2. Concentration of the hamming distance

Theorem 1 (Concentration). *For any $\varepsilon > 0$, the empirical Hamming distance concentrates around its expectation as*

$$\Pr\left(\left|d_H(P_i, P_j) - \frac{\theta_{ij}}{\pi}\right| \geq \varepsilon\right) \leq 2\exp(-2\, d\varepsilon^2). \tag{C.3}$$

Proof. Each bit disagreement in the Hamming distance is an independent Bernoulli trial with success probability θ_{ij}/π. The normalized Hamming distance is therefore the empirical mean of d i.i.d. Bernoulli variables. Applying Hoeffding's inequality directly yields the stated bound. □

C.3. Rationale for dimension selection

The projection dimension d governs both the stability and the capacity of the Hamming space. While signed random projection is inherently probabilistic, increasing d reduces the variance of the angular distance estimation (as per Theorem 1) and lowers the probability of code collisions. In practice, for N semantic categories, we choose

$$d = \lceil \log_2 N + d_{\text{free}} \rceil, \tag{C.4}$$

where d_{free} is a small safety margin. As shown in Fig. 4, performance saturates once d is sufficiently large to produce stable Hamming representations, indicating diminishing returns from further increasing the dimensionality.

References

[1] Y. Liu, Y. Dang, X. Gao, J. Han, L. Shao, Zero-shot sketch-based image retrieval via adaptive relation-aware metric learning, Pattern Recognit. 152 (2024) 110452.
[2] Y. Li, S. Deng, C. Guan, J. Gao, Complementary two-branch transformer for multi-label image retrieval, Pattern Recognit. 168 (2025) 111806.
[3] S. Chang, P. Ghamisi, Changes to captions: an attentive network for remote sensing change captioning, IEEE Trans. Image Process. 32 (2023) 6047–6060.
[4] M. Sun, W. Wang, X. Zhu, J. Liu, Reparameterizing and dynamically quantizing image features for image generation, Pattern Recognit. 146 (2024) 109962.
[5] Y. Xu, W. Yu, P. Ghamisi, M. Kopp, S. Hochreiter, Txt2Img-MHN: remote sensing image generation from text using modern Hopfield networks, IEEE Trans. Image Process. 32 (2023) 5737–5750.
[6] Z. Xiao, G. Diao, Z. Deng, Fine grained food image recognition based on swin transformer, J. Food Eng. 380 (2024) 112134.
[7] Y. Liu, W. Min, S. Jiang, Y. Rui, Convolution-enhanced bi-branch adaptive transformer with cross-task interaction for food category and ingredient recognition, IEEE Trans. Image Process. 33 (2024) 2572–2586.
[8] P. Zhou, W. Min, J. Song, Y. Zhang, S. Jiang, Synthesizing knowledge-enhanced features for real-world zero-shot food detection, IEEE Trans. Image Process. 33 (2024) 1285–1298.
[9] G. Qiao, D. Zhang, N. Zhang, X. Shen, X. Jiao, W. Lu, D. Fan, J. Zhao, H. Zhang, W. Chen, J. Zhu, Food recommendation towards personalized wellbeing, Trends Food Sci. Technol. 156 (2025) 104877.
[10] W. Min, Z. Wang, Y. Liu, M. Luo, L. Kang, X. Wei, X. Wei, S. Jiang, Large scale visual food recognition, IEEE Trans. Pattern Anal. Mach. Intell. 45 (8) (2023) 9932–9949.
[11] A.-K.N. Vu, T.-T. Do, N.-D. Nguyen, V.-T. Nguyen, T.D. Ngo, T.V. Nguyen, Instance-level few-shot learning with class hierarchy mining, IEEE Trans. Image Process. 32 (2023) 2374–2385.
[12] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[13] K. Zhou, J. Yang, C.C. Loy, Z. Liu, Learning to prompt for vision-language models, Int. J. Comput. Vis. 130 (9) (2022) 2337–2348.
[14] K. Zhou, J. Yang, C.C. Loy, Z. Liu, Conditional prompt learning for vision-language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16816–16825.
[15] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, T. Duerig, Scaling up visual and vision-language representation learning with noisy text supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 4904–4916.
[16] J. Li, D. Li, C. Xiong, S. Hoi, Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International Conference on Machine Learning, PMLR, 2022, pp. 12888–12900.
[17] A. Baldrati, M. Bertini, T. Uricchio, A. Del Bimbo, Effective conditioned and composed image retrieval combining clip-based features, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 21466–21474.
[18] A. Sain, A.K. Bhunia, P.N. Chowdhury, S. Koley, T. Xiang, Y.-Z. Song, Clip for all things zero-shot sketch-based image retrieval, fine-grained or not, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2765–2775.
[19] Z. Wang, Y. Lu, Q. Li, X. Tao, Y. Guo, M. Gong, T. Liu, Cris: clip-driven referring image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11686–11695.
[20] T. Lüddecke, A. Ecker, Image segmentation using text and image prompts, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7086–7096.
[21] S. Eslami, C. Meinel, G. De Melo, Pubmedclip: how much does clip benefit visual question answering in the medical domain?, in: Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 1181–1193.
[22] M.U. Khattak, H. Rasheed, M. Maaz, S. Khan, F.S. Khan, Maple: multi-modal prompt learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19113–19122.
[23] M.U. Khattak, S.T. Wasim, M. Naseer, S. Khan, M.-H. Yang, F.S. Khan, Self-regulating prompts: foundational model adaptation without forgetting, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15190–15200.
[24] T. Zheng, Z.-D. Chen, Z.-C. Zhang, Z.-X. Ma, L.-J. Zhao, C.-Y. Zhang, X. Luo, X.-S. Xu, DGPrompt: Dual-guidance prompts generation for vision-language models, Neural Networks (2025) 107472.
[25] Y. Lu, J. Liu, Y. Zhang, Y. Liu, X. Tian, Prompt distribution learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5206–5215.
[26] H. Yao, R. Zhang, C. Xu, Visual-language prompt tuning with knowledge-guided context optimization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6757–6767.
[27] H. Yao, R. Zhang, C. Xu, TCP: textual-based class-aware prompt tuning for visual-language model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 23438–23448.
[28] L. Zhou, M. Ye, S. Li, N. Li, X. Zhu, L. Deng, H. Liu, Z. Lei, Bayesian test-time adaptation for vision-language models, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 29999–30009.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, in: Proceedings of the International Conference on Learning Representations, 2021.
[30] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[31] B. Zhang, P. Zhang, X. Dong, Y. Zang, J. Wang, Long-clip: unlocking the long-text capability of clip, in: European Conference on Computer Vision, Springer, 2024, pp. 310–325.
[32] W.B. Johnson, J. Lindenstrauss, et al., Extensions of Lipschitz mappings into a Hilbert space, Contemp. Math. 26 (189–206) (1984) 1.
[33] P. Yaskov, A short proof of the Marchenko–Pastur theorem, C.R. Math. 354 (3) (2016) 319–322.
[34] D.-N. Zou, S.-H. Zhang, T.-J. Mu, M. Zhang, A new dataset of dog breed images and a benchmark for finegrained classification, Comput. Visual Media 6 (2020) 477–487.
[35] O.M. Parkhi, A. Vedaldi, A. Zisserman, C.V. Jawahar, Cats and dogs, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3498–3505.
[36] M.-E. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, IEEE, 2008, pp. 722–729.
[37] J. Krause, M. Stark, J. Deng, L. Fei-Fei, 3D object representations for fine-grained categorization, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013, pp. 554–561.
[38] Z. Sun, Y. Yao, X.-S. Wei, Y. Zhang, F. Shen, J. Wu, J. Zhang, H.T. Shen, Webly supervised fine-grained recognition: benchmark datasets and an approach, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10602–10611.
[39] A. Khosla, N. Jayadevaprakash, B. Yao, L. Fei-Fei, Novel dataset for fine-grained image categorization, in: First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, 2011.
[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: CVPR09, 2009.
[41] S. Hou, Y. Feng, Z. Wang, Vegfru: a domain-specific dataset for fine-grained visual categorization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 541–549.
[42] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The Caltech-UCSD Birds-200-2011 Dataset, 2011, (California Institute of Technology).
[43] L. Bossard, M. Guillaumin, L. Van Gool, Food-101–mining discriminative components with random forests, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI 13, Springer, 2014, pp. 446–461.
[44] J. Chen, C.-W. Ngo, Deep-based ingredient recognition for cooking recipe retrieval, in: Proceedings of the 24th ACM International Conference on Multimedia, 2016, pp. 32–41.
[45] W. Min, L. Liu, Z. Luo, S. Jiang, Ingredient-guided cascaded multi-attention network for food recognition, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1331–1339.
[46] W. Min, L. Liu, Z. Wang, Z. Luo, X. Wei, X. Wei, S. Jiang, Isia food-500: a dataset for large-scale food recognition via stacked global-local attention network, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 393–401.
[47] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, A. Vedaldi, Fine-Grained Visual Classification of Aircraft, 2013, arXiv:1306.5151
[48] M.U. Khattak, M.F. Naeem, M. Naseer, L. Van Gool, F. Tombari, Learning to prompt with text only supervision for vision-language models, in: Proceedings of the AAAI Conference on Artificial Intelligence, 39, 2025, pp. 4230–4238.
[49] Z. Li, Y. Song, M.-M. Cheng, X. Li, J. Yang, Advancing textual prompt learning with anchored attributes, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 3618–3627.
[50] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[51] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, S.-N. Lim, Visual prompt tuning, in: European Conference on Computer Vision, Springer, 2022, pp. 709–727.
[52] Z. Wang, Z. Zhang, S. Ebrahimi, R. Sun, H. Zhang, C.-Y. Lee, X. Ren, G. Su, V. Perot, J. Dy, et al., Dualprompt: complementary prompting for rehearsal-free continual learning, in: European Conference on Computer Vision, Springer, 2022, pp. 631–648.