# Advancing Textual Prompt Learning with Anchored Attributes

Supplementary Material

## 1. Implementation Details

### 1.1. Dataset

We evaluate the performance of our method on 15 recognition datasets. For generalization from base-to-novel classes and cross-dataset evaluation, we evaluate the performance of our method on 11 diverse recognition datasets. Specifically, these datasets include ImageNet-1K [3] and Caltech-101 [4] for generic object classification; OxfordPets [14], StanfordCars [9], Flowers-102 [13], Food-101 [1], and FGVCAircraft [12] for fine-grained classification, SUN-397 [19] for scene recognition, UCF-101 [17] for action recognition, DTD [2] for texture classification, and EuroSAT [5] for satellite imagery recognition. For domain generalization experiments, we use ImageNet-1K [3] as the source dataset and its four variants as target datasets including ImageNet-V2 [16], ImageNet-Sketch [18], ImageNet-A [7], and ImageNet-R [6].

### 1.2. Attribute Search

Inspired by DARTS [11], we employ a differentiable search method to identify the optimal content and quantity of attributes for our proposed attribute-anchored form. The search process is conducted for 10 epochs with a batch size of 32. We use SGD to optimize the soft prompts $\theta$ with an initial learning rate of $0.002$. and Adam to optimize the weight vector $\alpha$ with an initial learning rate of $0.02$. In our experiments, we use 5 attribute bases, which generate 31 (i.e., $C_5^1 + C_5^2 + C_5^3 + C_5^4 + C_5^5$) candidate combinations for the search process.

Tab. S4 presents the five attribute bases generated by the LLM, alongside the optimal attribute combination identified after the search. Furthermore, Tab. S5 displays the final weights of all candidate combinations from the search stage on the Caltech-101 dataset.

### 1.3. Base-to-Novel Generalization

**Baseline Methods.** To evaluate ATPrompt, we integrate it with several leading textual-based prompt learning approaches, including CoOp [22], CoCoOp [21], MaPLe [8], DePT [20] and PromptKD [10]. The experimental settings are detailed below.

**Settings.** Our framework is implemented in PyTorch [15] and all experiments were conducted on a single NVIDIA A800 GPU. Following the baseline methods, we use a standard data augmentation scheme of random resized cropping and flipping. We employe Stochastic Gradient Descent (SGD) as the optimizer. By default, the soft token lengths for attribute and class tokens are set to be identical, as attribute and class tokens are considered equally important. The specific implementation details for each baseline method are presented as follows:

**CoOp+ATPrompt:** Following the baseline, we use a batch size of 32 and an initial learning rate of 0.002. The original paper reports a learnable prompt length of $M = 16$ for ResNet-50 but does not specify a length for ViT-B/16. In our setup, we set the sofo token length for both the attribute and class tokens to 2. While the baseline model is trained for 200 epochs, we reduce the training to 100 epochs while maintaining the same cosine decay schedule. Figure S1 illustrates the architectural differences between the original CoOp and CoOp+ATPrompt.

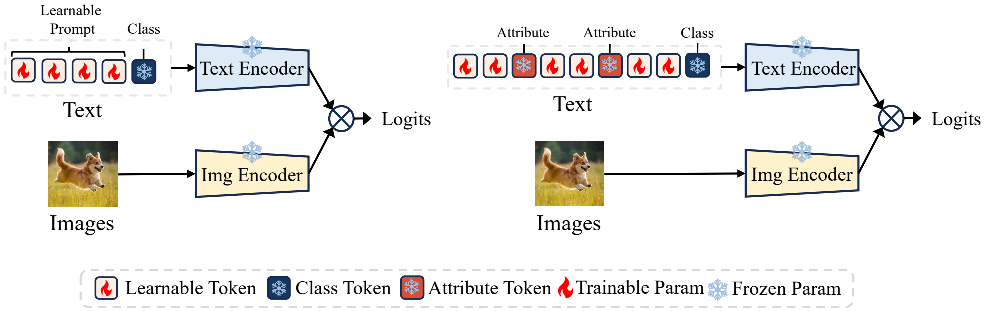
**CoCoOp+ATPrompt**: We adhere to the baseline's settings with a batch size of 1 and an initial learning rate of 0.002. Whereas the original paper specifies a soft class token length of 4, we set the length of our learnable tokens for the attribute and class token to 2. We adopt the same training schedule as the baseline: 10 epochs with cosine decay.

CoCoOp's original design uses a meta-network to generate offsets for all soft prompt tokens. We retain this meta-network but modify its application: the meta tokens now serve as offsets exclusively for the class soft tokens, $[T_1]$, ..., $[T_M]$, as shown in Fig. S2.

**MaPLe+ATPrompt**: We adhere to the baseline hyperparameters, utilizing a batch size of 4 and an initial learning rate of 0.0035. We diverge from the original prompt configuration; whereas the baseline sets the learnable prompt length to 2, our method sets the soft token lengths of both the attribute and the class token to 4. The training schedule remains consistent with the baseline.

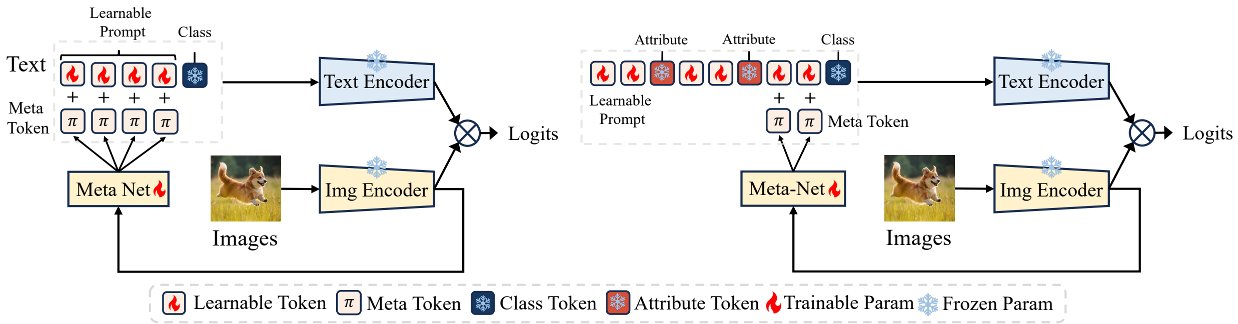
The primary architectural modification in MaPLe + ATPrompt concerns the projection mechanism. The original MaPLe framework inputs all textual soft tokens into a projection layer to generate corresponding visual tokens, which are then fused into the image encoder. Our approach, however, selectively inputs only the class soft tokens into this projection layer, while the attribute tokens are preserved without modification. This architectural difference is visualized in Fig. S3.

**DePT+ATPrompt:** We adopt the baseline's training configuration: a batch size of 32, a learning rate of 0.0035, a balance weight of $\lambda$=0.7, and a duration of 10 epochs. Our primary configuration for DePT+ATPrompt uses a learnable token length of 4. For datasets with lower complexity, namely Caltech-101, OxfordPets, and StanfordCars, we adjust these parameters, setting the soft token length to 2 and the balance weight to 0.6. The architectural differences between the DePT and DePT+ATPrompt models are detailed



(a) CoOp (b) CoOp + ATPrompt

Figure S1. Architectural comparison between CoOp and CoOp+ATPrompt.



(a) CoCoOp (b) CoCoOp + ATPrompt

Figure S2. Architectural comparison between CoCoOp and CoCoOp+ATPrompt. In CoCoOp+ATPrompt, meta tokens are only added as offsets to class soft tokens.

in Fig. S4.

## 2. Additional Experiments

### 2.1. Ablation Study

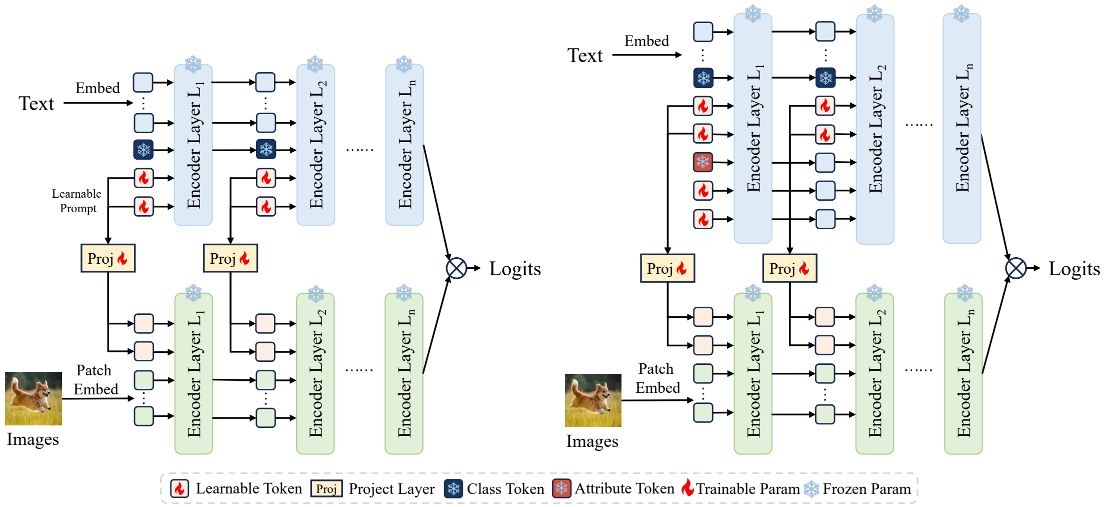
**Attribute Order.** In the main paper, our experiments confirm that the order of attributes does not significantly impact model performance, with results fluctuating within an acceptable range. Here we provide additional experiments in Tab. S1 to support this observation.

**Attribute Position.** We also investigated the impact of attribute token positioning within the prompt. Fig. S5 visualizes the positions tested, and Tab. S2 presents the results. Our findings show that the "interval" configuration, where attributes are placed between class tokens, yields the best performance.

**Initialization.** Baseline methods typically initialize soft tokens using the embeddings of the phrase "a photo of a." The inclusion of attribute tokens makes this strategy suboptimal for our method. We instead initialize class soft tokens ($[T_1], ..., [T_M]$) by sampling from a Gaussian distribution with a mean of 0 and a standard deviation of 0.02. As shown in Table S3, this random initialization provides a superior starting point for training.

| Attributes | Base | Novel | HM |
|---|---|---|---|
| (shape, color) | 76.32 | 70.39 | 73.24 |
| (color, shape) | 76.27 | 70.60 | **73.33** |
| (size, habitat) | 76.44 | 70.23 | 73.20 |
| (habitat, size) | 76.46 | 70.16 | 73.14 |
| (material, function) | 76.40 | 70.13 | 73.13 |
| (function, material) | 76.28 | 70.00 | 73.01 |
| (growth, season) | 76.46 | 70.18 | 73.19 |
| (season, growth) | 76.40 | 70.21 | 73.17 |
| (color, size, shape) | 76.27 | 69.95 | 72.97 |
| (shape, size, color) | 76.32 | 70.19 | 73.13 |
| (habitat, size, shape) | 76.50 | 70.21 | 73.22 |
| (habitat, shape, size) | 76.46 | 70.08 | 73.13 |
| Searched Attributes<br>(color, shape) | 76.27 | 70.60 | **73.33** |

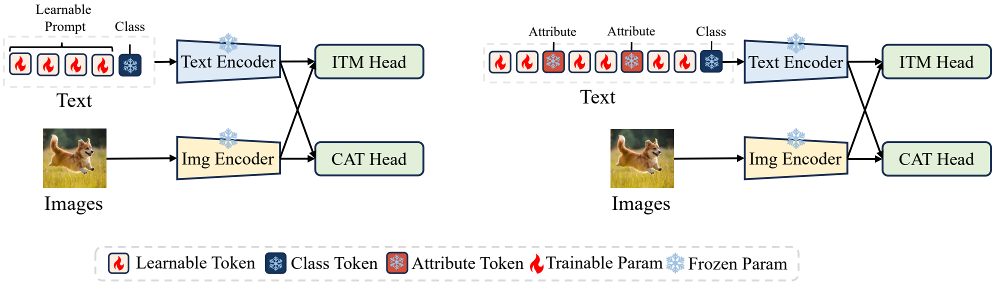
Table S1. Comparison of different attribute orders on ImageNet. Changes in attribute order will not significantly affect model performance.



(a) MaPLe

(b) MaPLe + ATPrompt

Figure S3. Architectural comparison between MaPLe and MaPLe+ATPrompt.



(a) DePT

(b) DePT + ATPrompt

Figure S4. Architectural comparison between DePT and DePT+ATPrompt.

| Version | Base | Novel | HM |
|---|---|---|---|
| Baseline (CoOp) | 76.47 | 67.88 | 71.92 |
| (a) Interval (Ours) | 76.27 | 70.60 | 73.33 |
| (b) Adjacent-front | 76.39 | 70.22 | 73.18 |
| (c) Adjacent-middle | 76.46 | 70.11 | 73.15 |
| (d) Adjacent-end | 76.34 | 70.31 | 73.20 |
| (e) Separate | 76.48 | 70.08 | 73.14 |

Table S2. Performance results of attribute tokens at different positions in ATPrompt on ImageNet. The interval version achieves best results.

| Attribute | Base | Novel | HM |
|---|---|---|---|
| “a photo of a” | 76.40 | 70.07 | 73.10 |
| Random Normal Init | 76.27 | 70.60 | 73.33 |

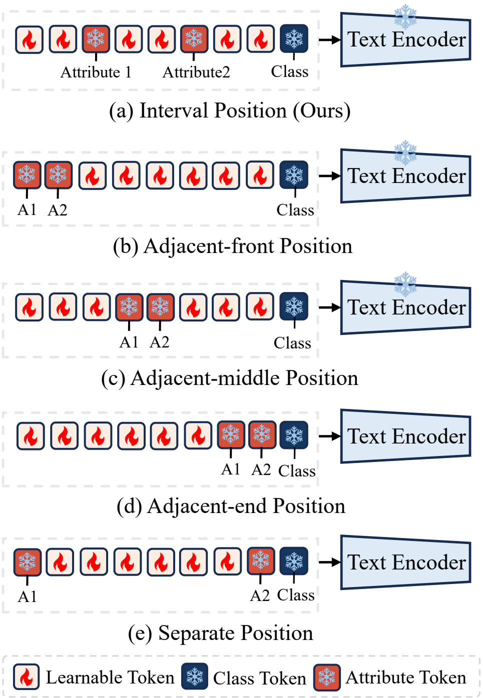
Table S3. Comparison of different initialization ways on ImageNet. Random normal initialization performs better.

## 3. Discussion

**Comparison with Direct LLM Queries.** Directly querying an LLM for universal attributes presents two challenges: determining the optimal attribute content and identifying the ideal number of attributes. Our experiments suggest that two attributes are often optimal. Therefore, users can bypass our search process by directly prompting the LLM to summarize two universal attributes. This offers a simpler approach, though it may result in a slight performance trade-off.

**Why do attributes searched on a source dataset (ImageNet) generalize well?** The attributes identified on ImageNet (e.g., color, shape) are fundamental properties of natural objects. Representations learned under the guidance of these universal attributes are therefore inherently generalizable and transfer effectively to other datasets and classes.

**Why does ATPrompt not outperform regularization-based methods in isolation?** ATPrompt is a plug-in module designed to optimize the prompt’s structure. In contrast, regularization-based methods are often comprehensive frameworks that employ multiple components (e.g., learnable visual prompts, MLPs) simultaneously. While ATPrompt may not outperform these multi-faceted approaches on its own, its strength lies in its ability to be



(a) Interval Position (Ours)

(b) Adjacent-front Position

(c) Adjacent-middle Position

(d) Adjacent-end Position

(e) Separate Position

Figure S5. Comparison of attribute tokens at different positions, taking two attributes as an example.

integrated into other methods, consistently improving their performance beyond previous baselines.

## 4. Limitations and Future Works.

Beyond the limitations discussed in the main paper, we identify the following directions for future research: (1) While our differentiable search method is efficient, we aim to further enhance the attribute discovery process. A promising direction is to leverage Multimodal Large Language Models (MLLMs), potentially using techniques like Chain-of-Thought (CoT), to better automate the selection of optimal attribute content and quantity. (2) Our current approach embeds fixed, explicit attributes into the prompt. In the future, we plan to explore a transition to implicit, learnable attributes. This would enable the model to discover optimal attributes in a data-driven manner during training, potentially unlocking further performance gains.

| Dataset | Attribute Bases | Searched Attributes |
|---|---|---|
| ImageNet-1K | color, size, shape, habitat, behavior | (color, shape) |
| Caltech-101 | shape, color, material, function, size | (shape,size) |
| Oxford Pets | loyalty, affection, playfulness, energy, intelligence | (playfulness, energy) |
| Stanford Cars | design, engine, performance, luxury, color | (luxury) |
| Flowers-102 | color, flower, habitat, growth, season | (color, habitat, growth) |
| Food-101 | flavor, texture, origin, ingredients, preparation | (flavor, preparation) |
| FGVC Aircraft | design, capacity, range, engines, liveries | (design, range) |
| SUN-397 | architecture, environment, structure, design, function | (function) |
| DTD | pattern, texture, color, design, structure | (pattern, color, design) |
| EuroSAT | habitat, foliage, infrastructure, terrain, watercourse | (habitat) |
| UCF-101 | precision, coordination, technique, strength, control | (precision) |

Table S4. Attribute bases and searched results for each dataset.

| Attribute Bases | shape, color, material, function, size |
|---|---|
| Attribute Combinations & Corresponding Weights | (shape), weight: 0.298 |
| | (color), weight: 0.004 |
| | (material), weight: 0.002 |
| | (function), weight: 0.002 |
| | (size), weight: 0.003 |
| | (shape, color), weight: 0.003 |
| | (shape, material), weight: 0.006 |
| | (shape, function), weight: 0.000 |
| | **(shape, size), weight: 0.565** |
| | (color, material), weight: 0.000 |
| | (color, function), weight: 0.001 |
| | (color, size), weight: 0.005 |
| | (material, function), weight: 0.000 |
| | (material, size), weight: 0.002 |
| | (function, size), weight: 0.002 |
| | (shape, color, material), weight: 0.002 |
| | (shape, color, function), weight: 0.002 |
| | (shape, color, size), weight: 0.000 |
| | (shape, material, function), weight: 0.001 |
| | (shape, material, size), weight: 0.085 |
| | (shape, function, size), weight: 0.001 |
| | (color, material, function), weight: 0.001 |
| | (color, material, size), weight: 0.000 |
| | (color, function, size), weight: 0.002 |
| | (material, function, size), weight: 0.001 |
| | (shape, color, material, function), weight: 0.001 |
| | (shape, color, material, size), weight: 0.001 |
| | (shape, color, function, size), weight: 0.001 |
| | (shape, material, function, size), weight: 0.005 |
| | (color, material, function, size), weight: 0.001 |
| | (shape, color, material, function, size), weight: 0.001 |

Table S5. Output results after 40 epochs of attribute searching on the Caltech101 dataset.

# References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461. Springer, 2014. 1

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1

[4] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR workshop*, pages 178–178. IEEE, 2004. 1

[5] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 1

[6] 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 *ICCV*, pages 8340–8349, 2021. 1

[7] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021. 1

[8] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023. 1

[9] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshop*, pages 554–561, 2013. 1

[10] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *CVPR*, pages 26617–26626, 2024. 1

[11] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018. 1

[12] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1

[13] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 1

[14] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE, 2012. 1

[15] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8026–8037, 2019. 1

[16] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400. PMLR, 2019. 1

[17] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[18] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 32, 2019. 1

[19] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492. IEEE, 2010. 1

[20] Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. In *CVPR*, pages 12924–12933, 2024. 1

[21] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 1

[22] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 1