

Causality-guided Prompt Learning for Vision-language Models via Visual Granulation

Mengyu Gao¹² Qiulei Dong¹² *

¹ State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences

² School of Artificial Intelligence, University of Chinese Academy of Sciences

gaomengyu2021@ia.ac.cn

qldong@nlpr.ia.ac.cn.

Abstract

Prompt learning has recently attracted much attention for adapting pre-trained vision-language models (e.g., CLIP) to downstream recognition tasks. However, most of the existing CLIP-based prompt learning methods only show a limited ability for handling fine-grained datasets. To address this issue, we propose a causality-guided text prompt learning method via visual granulation for CLIP, called CaPL, where the explored visual granulation technique could construct sets of visual granules for the text prompt to capture subtle discrepancies among different fine-grained classes through casual inference. The CaPL method contains the following two modules: (1) An attribute disentanglement module is proposed to decompose visual features into non-individualized attributes (shared by some classes) and individualized attributes (specific to single classes) using a Brownian Bridge Diffusion Model; (2) A granule learning module is proposed to construct visual granules by integrating the aforementioned attributes for recognition under two causal inference strategies. Thanks to the learned visual granules, more discriminative text prompt is expected to be learned. Extensive experimental results on 15 datasets demonstrate that our CaPL method significantly outperforms the state-of-the-art prompt learning methods, especially on fine-grained datasets. Code is available at https://github.com/GaoMY-521/CaPL_Code.

1. Introduction

CLIP[39], a typical vision-language model pre-trained on large-scale image-text pairs, has shown its zero-shot recognition ability in many existing works[12, 38, 41, 64]. Recently, prompt learning methods have attracted growing interests for further enhancing CLIP’s performance on downstream recognition tasks[13, 31, 36, 61].

*Corresponding author

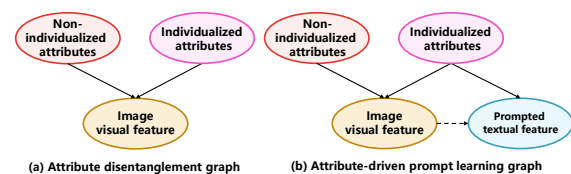


Figure 1. Causal graphs of (a) attribute disentanglement and (b) attribute-driven prompt learning for recognition.

In general, the existing prompt learning methods refine the inputs of CLIP[39] with learnable prompts[14, 20, 26, 42], so that the visual and textual features outputted from CLIP are aligned more effectively. The early global prompt learning methods[52, 54, 55, 62] process the features holistically during prompt learning. Despite achieving improvements on coarse-grained recognition, they fall short in fine-grained recognition, since the subtle discrepancies among fine-grained classes cannot be captured by global features as indicated in [21, 58]. Recently, local prompt learning methods are proposed to focus on specific attributes, either identifying the most discriminative attributes while discarding others[28, 30, 48], or treating all attributes equally to encode attribute-specific representations[57–59]. However, these methods cannot allocate dynamic attention to all attributes as indicated in [11, 63], still leading to a limited performance on challenging fine-grained datasets (e.g., Flowers102[34] for flower recognition and FGVC Aircraft[33] for airplane recognition).

Unlike the existing methods that either discard non-discriminative attributes or handle all attributes uniformly, our goal is to disentangle visual features into individualized attributes (exclusive to single classes) and non-individualized attributes (shared by some classes), and perceive all attributes according to their discrimination ability, assuming that all attributes contribute to prompt learning but to varying degrees. This idea is supported by the following fact: In the fine-grained dataset Flowers102[34], petal color

has weak discrimination ability, since more than 20 classes share yellow petals, yet it still aids in distinguishing classes with pink and yellow petals, whereas petal shape, such as trumpet-shaped or heart-shaped, is strongly discriminative as it varies across classes. To model these relationships, we introduce the attribute disentanglement graph shown in Fig. 1(a) to capture the connections among visual features and attributes. Based on this, we propose the attribute-driven prompt learning graph shown in Fig. 1(b), which structures the relationship between a text prompt, visual features, and attributes. This graph could serve as a guidance for designing corresponding causal inference strategies that enable differentiated attribute perception for prompt learning.

Accordingly, we propose a causality-guided text prompt learning method, called CaPL, where attributes are disentangled from visual features to enable adaptive understanding through a visual granulation technique. Specifically, an attribute disentanglement module is firstly proposed to decompose the visual features extracted by CLIP[39] into non-individualized and individualized attribute representations, optimized within a Brownian Bridge Diffusion Model (BDDM)[29]-based network. Then, a granule learning module is proposed to construct sets of visual granules by integrating the disentangled attributes for recognition, where two causal inference strategies are applied: (1) For each of the individualized attributes, the factual intervention is to construct a corresponding factual granule by decorating this individualized attribute with all the non-individualized attributes; (2) To improve the generalizability of the text prompt, the counterfactual intervention is to construct counterfactual granules by swapping non-individualized and individualized attributes across different images. Finally, the text prompt is learned under the supervision of these visual granules during the training process.

Our contributions are summarized as follows:

- We construct an attribute-driven prompt learning graph, which could depict the relationship between a text prompt, visual features and attributes. Accordingly, we propose an attribute disentanglement module to disentangle attributes with different discrimination ability;
- We explore the visual granulation technique in the proposed granule learning module under two causal inference strategies. The text prompt learned by utilizing the constructed visual granules as supervision signals could capture fine-grained discrepancies among different classes as demonstrated in Sec. 4.3;
- By integrating the above modules, the causality-guided text prompt learning method is proposed, whose superiority is demonstrated in Sec. 4.2.

2. Related Work

Global prompt learning. These methods learn prompts by utilizing features extracted by CLIP[39] holistically. Some

of them project visual features to learn text prompts[6, 7, 14, 23, 35, 36, 49, 54, 55, 62]. Wang *et al.*[49] encoded visual features as a local bias of text prompts. Yao *et al.*[55] incorporated prior knowledge into class-aware text prompts. Recently, some methods project textual features to learn visual prompts[3, 22, 42, 51, 56]. Jia *et al.*[22] introduced visual prompt before each image encoder layer. Shi *et al.*[42] synthesized class name images as visual prompts. More recently, some researchers use both text and visual prompts for dual-modal feature alignment[13, 20, 24, 25, 48, 52]. Khattak *et al.*[24] injected dual-modal prompts into each CLIP encoder layer for mutual synergy. Hu *et al.*[20] regularized dual-modal prompts to retain pre-trained knowledge. However, these methods generally treat the visual and textual features as a whole, failing to capture differences between fine-grained classes and resulting in limited fine-grained recognition [21, 57].

Local prompt learning. These methods extract attribute-specific features to capture subtle differences between fine-grained classes[28, 30, 48, 57–59, 63]. Lafon *et al.*[28] selected top-k discriminative attributes for prompt learning. Wang *et al.*[48] proposed hierarchical tuning to preserve only discriminative attributes in visual features. Zhang *et al.* extracted concept-level visual features describing attributes by a visual concept cache[58] or a conceptual codebook[57]. Zhang *et al.*[59] decoupled semantics from visual features to learn text prompts gradually. Despite their advancement in fine-grained recognition, they still fall short on challenging fine-grained datasets since they ignore that attributes contribute differently to recognition [11, 63]. Unlikely, we propose a causality-guided text prompt learning method via visual granulation, enabling perception of attributes based on their discrimination ability.

3. Methodology

3.1. Architecture

Fig. 2 shows the architecture of the proposed causality-guided text prompt learning method, which contains a pre-trained CLIP consisting of an image encoder and a text encoder, an attribute disentanglement module, and a granule learning module to learn a text prompt.

As shown in Fig. 2(a), at the training stage, the image encoder is used to extract visual features from images, and the text encoder is used to extract prompted textual features from a combination of a learnable text prompt and class names following[54, 61]. Then, the attribute disentanglement module decomposes the non-individualized and individualized attribute representations from visual features. According to the attribute-driven prompt learning graph in Fig. 1(b), the granule learning module uses both the non-individualized and individualized attributes disentangled from the attribute disentanglement module to construct

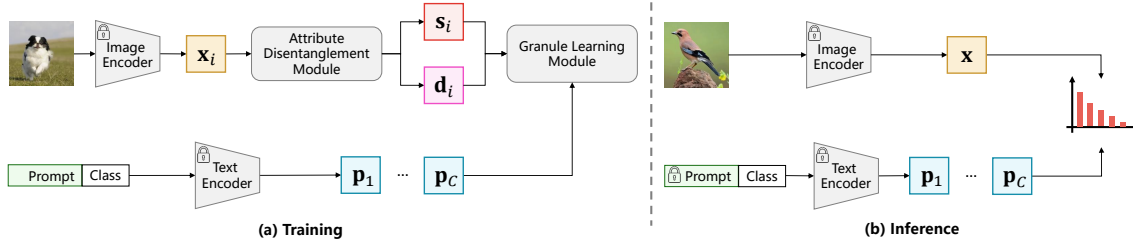


Figure 2. Architecture of CaPL, where (a) is the training stage and (b) is the inference stage. \mathbf{x}_i and \mathbf{x} are visual features, \mathbf{s}_i and \mathbf{d}_i are the non-individualized and individualized attribute representations, $\mathbf{p}_1, \dots, \mathbf{p}_C$ are the prompted textual features generated from a learnable text prompt and class names, C is the number of classes, and the “lock” symbol denotes the corresponding parameters are fixed.

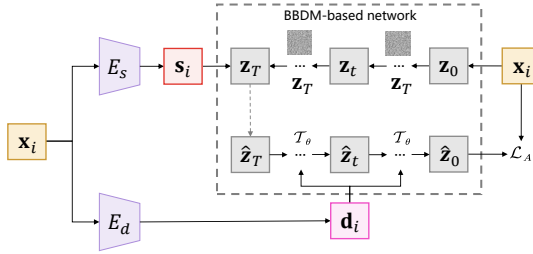


Figure 3. Architecture of attribute disentanglement module, which contains two encoders E_s, E_d to extract non-individualized and individualized attribute representation $\mathbf{s}_i, \mathbf{d}_i$ from the visual feature \mathbf{x}_i respectively, and a BBDM-based network. The upper feature transfer process of BBDM is the diffusion process, which generates latent features $\mathbf{z}_0, \dots, \mathbf{z}_T$. The lower one is the reverse process, which generate reconstructed features $\hat{\mathbf{z}}_T, \dots, \hat{\mathbf{z}}_0$ gradually, \mathcal{T}_θ is a learnable transfer model, and \mathcal{L}_A is the training loss.

visual granules, which are served as supervision signals for the text prompt to perceive specific attributes.

As shown in Fig. 2(b), at the inference stage, a testing image is recognized by comparing the cosine similarity between its visual feature and all prompted textual features.

3.2. Attribute Disentanglement Module

The attribute disentanglement module is proposed to decompose visual features into non-individualized and individualized attribute representations. In order to model the transition relationship among non-individualized attribute representations, individualized attribute representations, and the input visual features, a BBDM (Brownian Bridge Diffusion Model)[29] is introduced in this module as shown in Fig. 3, inspired by the success of BBDM in modeling the relationship among different feature distributions in other tasks [32, 47], and ability of diffusion process to provide inductive biases for feature disentanglement[53].

Specifically, given the i -th image from class c_i , its visual feature \mathbf{x}_i is extracted from the frozen CLIP image encoder. Two encoders E_s and E_d are used to obtain non-individualized attribute representation \mathbf{s}_i and individualized

attribute representation \mathbf{d}_i respectively.

Then, the BBDM is used to optimize the disentangled attributes by learning to reverse the disentanglement process. It is noted that the input visual features are regarded as target since they integrate both non-individualized and individualized attributes, while the individualized attribute representations serve as conditions due to their class-specific information, in contrast to the non-individualized ones that capture shared attributes. Therefore, we model the feature transition from non-individualized attribute representation \mathbf{s}_i to the input visual feature \mathbf{x}_i conditioned on the individualized attribute representation \mathbf{d}_i . Following [29], the diffusion process transfers $\mathbf{z}_0 = \mathbf{x}_i$ to $\mathbf{z}_T = \mathbf{s}_i$ following the Brownian Bridge stochastic process:

$$q(\mathbf{z}_t | \mathbf{z}_0, \mathbf{z}_T) = \mathcal{N}(\mathbf{z}_t; (1 - m_t)\mathbf{z}_0 + m_t\mathbf{z}_T, \delta_t \mathbf{I}) \quad (1)$$

where $t \in [1, T]$, T is the transfer step, \mathbf{z}_t is a latent feature, $m_t = t/T$, $\delta_t = 2(m_t - m_t^2)$, $\mathcal{N}(\mathbf{z}; \mu, \Sigma)$ is a Gaussian distribution with mean μ , variance Σ , \mathbf{I} is an identity matrix. The reverse process learns to reverse the diffusion process by transferring $\hat{\mathbf{z}}_T = \mathbf{s}_i$ to $\hat{\mathbf{z}}_0 = \hat{\mathbf{x}}_i$ conditioned on \mathbf{d}_i :

$$p_\theta(\hat{\mathbf{z}}_{0:T}) = p(\hat{\mathbf{z}}_T) \prod_{t=1}^T p_\theta(\hat{\mathbf{z}}_{t-1} | \hat{\mathbf{z}}_t, \hat{\mathbf{z}}_T, \mathbf{d}_i) \quad (2)$$

where $p_\theta(\hat{\mathbf{z}}_{t-1} | \hat{\mathbf{z}}_t, \hat{\mathbf{z}}_T, \mathbf{d}_i)$ follows a Gaussian distribution with mean $\mu_\theta(\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_T, \mathbf{d}_i)$ and variance $\sigma_t^2 = \frac{\delta_{t-1}}{\delta_t} \left[\delta_t - \delta_{t-1} \frac{(1-m_t)^2}{(1-m_{t-1})^2} \right]$, $\hat{\mathbf{z}}_t$ is the reconstructed feature at time step t . The mean $\mu_\theta(\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_T, \mathbf{d}_i)$ is parameterized as a learnable transfer model $\mathcal{T}_\theta(\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_T, \mathbf{d}_i)$ to predict the difference $\Delta = \mathbf{z}_T - \mathbf{z}_0$. Following [29], the transfer model \mathcal{T}_θ is trained by reconstructing \mathbf{z}_0 directly from a random \mathbf{z}_t sampled from the diffusion process, rather than iterating T steps following the reverse process. With \mathbf{z}_t , the difference $\hat{\Delta}_t = \mathcal{T}_\theta(\mathbf{z}_t, \mathbf{z}_T, \mathbf{d}_i)$ is predicted to obtain the restored $\tilde{\mathbf{z}}_0 = \mathbf{z}_T - \hat{\Delta}_t$, and the training loss \mathcal{L}_A is to minimize the distance between restored $\tilde{\mathbf{x}}_i = \tilde{\mathbf{z}}_0$ and original $\mathbf{x}_i = \mathbf{z}_0$:

$$\mathcal{L}_A = \|\tilde{\mathbf{x}}_i - \mathbf{x}_i\|_2^2 \quad (3)$$

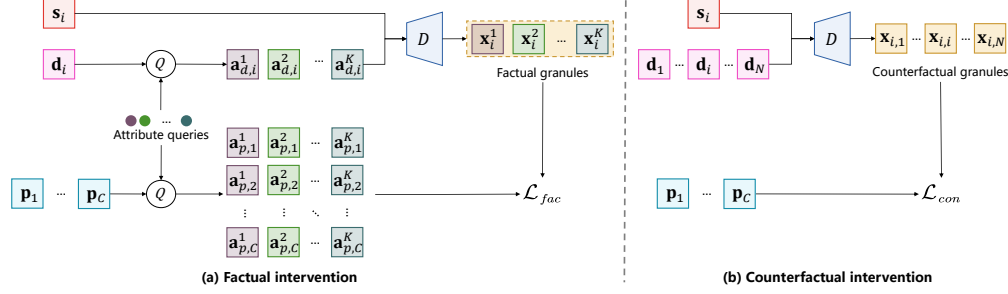


Figure 4. Architecture of granule learning module, which has two forms for (a) factual intervention and (b) counterfactual intervention. “ Q ” is the query process, $\{\mathbf{a}_{d,i}^k\}_{k=1}^K$ and $\{\mathbf{a}_{p,c}^k\}_{c=1, k=1}^{C,K}$ are the visual and textual representations of each individualized attribute, $\{\mathbf{d}_i\}_{i=1}^N$ are N individualized attributes in a training bath, D is the decoder to generate visual granules, \mathcal{L}_{fac} , \mathcal{L}_{con} are training losses.

After training BBDM, the visual features could be effectively decomposed into two representations capturing non-individualized and individualized attributes respectively.

3.3. Granule Learning Module

Notation. The text prompt is defined as M learnable context vectors $\mathbf{v} = [\mathbf{v}_1, \dots, \mathbf{v}_M]$ following [61]. The prompted textual feature \mathbf{p}_c of class c is extracted by the fixed CLIP text encoder from a concatenation of \mathbf{v} and \mathbf{e}_c , where \mathbf{e}_c is the class name embedding obtained by CLIP tokenizer.

Then, the granule learning module is proposed to utilize the visual granulation technique for learning the text prompt. According to the attribute-driven prompt learning graph in Fig. 1(b), attributes contribute differently to learning a discriminative text prompt. To explicitly model these differences, the visual granulation technique constructs sets of visual granules as supervision signals by integrating different attributes under two causal inference strategies:

Factual intervention. Since individualized attributes are highly discriminative while non-individualized attributes have relatively weak discrimination ability, the factual intervention is applied to construct factual granules, as shown in Fig. 4(a), by decorating each of the individualized attributes with all non-individualized attributes to provide a complete yet focused representation for recognition. By serving as supervision signals for text prompt learning, the effect of single individualized attributes could be emphasized to enhance fine-grained recognition.

Specifically, a series of learnable attribute queries $\mathbf{q}^1, \dots, \mathbf{q}^K$, each corresponding to an individualized attribute, is defined to disentangle the visual representations $\{\mathbf{a}_{d,i}^k\}_{k=1}^K$ and textual representations $\{\mathbf{a}_{p,c}^k\}_{c=1, k=1}^{C,K}$ of each individualized attribute from \mathbf{d}_i and $\mathbf{p}_1, \dots, \mathbf{p}_C$ respectively, where K is a preset number of individualized attributes. For the k -th individualized attribute, its visual representation $\mathbf{a}_{d,i}^k$ is obtained as follows:

$$\mathbf{a}_{d,i}^k = Q(\mathbf{d}_i, \mathbf{q}^k) = \mathcal{S}\left((\mathbf{q}^k)^T \times \mathbf{d}_i / \sqrt{d}\right) \times \mathbf{d}_i \quad (4)$$

where Q is the query process, $\mathcal{S}(\cdot)$ is the softmax function, $(\cdot)^T$ is the transpose operation, and d is the dimension of \mathbf{d}_i . The textual representation $\mathbf{a}_{p,c}^k$ of an arbitrary class c can be obtained by: $\mathbf{a}_{p,c}^k = Q(\mathbf{p}_c, \mathbf{q}^k)$ accordingly.

Then, the factual granule \mathbf{x}_i^k is constructed by combining $\mathbf{a}_{d,i}^k$ with the non-individualized representation \mathbf{s}_i : $\mathbf{x}_i^k = D(\mathbf{s}_i, \mathbf{a}_{d,i}^k)$, where D is a MLP-based decoder.

Finally, the textual representations are learned to recognize the factual granules. On the one hand, the textual representations $\{\mathbf{a}_{p,c}^k\}_{c=1, k=1}^{C,K}$ should recognize which individualized attribute constitutes the factual granule \mathbf{x}_i^k . The probability $p(y_a = k | \mathbf{x}_i^k)$ is defined as follows:

$$p(y_a = k | \mathbf{x}_i^k) = \frac{\exp(\cos(\mathbf{x}_i^k, \mathbf{a}_{p,c_i}^k) / \tau)}{\sum_{k'=1}^K (\exp(\cos(\mathbf{x}_i^k, \mathbf{a}_{p,c_i}^{k'}) / \tau))} \quad (5)$$

where y_a is the predicted individualized attribute, c_i is the class of \mathbf{x}_i , and τ is temperature coefficient.

On the other hand, the textual representations are learned to predict the class of \mathbf{x}_i by calculating the cosine similarity between K pairs of $\{\mathbf{x}_i^k\}_{k=1}^K$ and $\{\mathbf{a}_{p,c_i}^k\}_{k=1}^K$:

$$p(y_v = c_i | \mathbf{x}_i) = \frac{\sum_{k=1}^K \exp(\cos(\mathbf{x}_i^k, \mathbf{a}_{p,c_i}^k) / \tau)}{\sum_{c=1}^C \sum_{k=1}^K (\exp(\cos(\mathbf{x}_i^k, \mathbf{a}_{p,c}^k) / \tau))} \quad (6)$$

where y_v is the predicted class. The training loss \mathcal{L}_{fac} is the weighted sum of two cross entropy losses:

$$\mathcal{L}_{fac} = -\log(p(y_a = k | \mathbf{x}_i^k)) - \lambda_f \log(p(y_v = c_i | \mathbf{x}_i)) \quad (7)$$

where λ_f is a constant weight. Note: The ablation study by omitting the non-individualized attributes is in Sec. 4.3.

Counterfactual intervention. As stated in [1, 44], homogeneous non-individualized attributes may lead to spurious correlations in recognizing visual features, which reduces the generalizability of the text prompt when encountering

heterogeneous attributes. Therefore, the counterfactual intervention is applied to construct counterfactual granules, as shown in Fig. 4(b), by mixing non-individualized and individualized attributes across images to simulate alternative contexts. By serving as another supervision signals, the generalizability of the text prompt is strengthened.

For the i -th image, given its \mathbf{s}_i and all $\mathbf{d}_j, j = 1, \dots, N$ in a training batch, where N is batch size, the counterfactual granule $\mathbf{x}_{i,j}$ is obtained by: $\mathbf{x}_{i,j} = D(\mathbf{s}_i, \mathbf{d}_j)$. Then, the counterfactual granule $\mathbf{x}_{i,j}$ is assigned to the same class c_j of the j -th image, from which \mathbf{d}_j is decomposed. The probability $p(y_v = c_j | \mathbf{x}_{i,j})$ for recognizing $\mathbf{x}_{i,j}$ is based on the cosine similarity between $\mathbf{x}_{i,j}$ and $\mathbf{p}_1, \dots, \mathbf{p}_C$.

Furthermore, to guarantee that the decoder D could generate accurate visual granules, so that the recognition of $\mathbf{x}_{i,j}$ depends solely on the generalizability of the text prompt, we propose a reconstruction error to minimize the distance between $\mathbf{x}_{i,i} = D(\mathbf{s}_i, \mathbf{d}_i)$ and its corresponding input visual feature \mathbf{x}_i . The training loss \mathcal{L}_{con} is the weighed sum of the cross entropy loss and reconstruction error:

$$\mathcal{L}_{con} = -\log(p(y_v = c_j | \mathbf{x}_{i,j})) + \lambda_r \|\mathbf{x}_i - \mathbf{x}_{i,i}\|_2^2 \quad (8)$$

where λ_r is a constant weight. The total training objective \mathcal{L}_G of the granule learning module is defined as:

$$\mathcal{L}_G = \mathcal{L}_{fac} + \mathcal{L}_{con} \quad (9)$$

It is noted that we do not apply the reconstruction error to factual granules, since they are constructed by utilizing only one individualized attribute and are actually different from their corresponding input visual features.

3.4. Training and Inference

The proposed CaPL utilizes an iterative training scheme, where each iteration contains two learning stages:

First, the attribute disentanglement module is learned by minimizing \mathcal{L}_A in Eq. (3). Then, with fixed attribute disentanglement module, the text prompt and the granule learning module are learned by minimizing \mathcal{L}_G in Eq. (9).

At the inference stage, the learned text prompt is used to generate prompted textual features of all classes. Then, the cosine similarity between a testing image visual feature and all prompted textual features is calculated for recognition.

4. Experiments

4.1. Tasks and Implementation Details

Tasks. We evaluate the proposed CaPL in the following three downstream recognition tasks:

(1) **Base-to-new generalization:** As done in [57, 60], our method is evaluated on 11 image recognition datasets listed in Table 1, and the classes in each dataset are equally

divided into base and new classes. Our model is trained with 16-shot images on base classes, and tested on base and new classes. The evaluation metrics include the average per-class Top-1 accuracies on base classes “Base” and new classes “New”, and the harmonic mean $H = \frac{2 \times \text{Base} \times \text{New}}{\text{Base} + \text{New}}$.

(2) **Cross-dataset transfer:** As done in [55, 57], the 11 datasets used in the above task are also used here for testing the cross-dataset transfer ability of the proposed method. Our model is trained with 16-shot images from ImageNet-1K[9] and then tested on the other 10 datasets. The evaluation metric is the average per-class Top-1 accuracy.

(3) **Cross-domain generalization:** As done in [36, 57], our method is trained with 16-shot images from ImageNet-1K[9], and then tested on 4 variants listed in Table 3 for verifying its cross-domain generalizability. The evaluation metric is the average per-class Top-1 accuracy.

Implementation details. The image and text encoders of CLIP are ViT-B-32[2] and transformer[45], which are pre-trained and kept frozen. The two encoders E_s, E_d and decoder D are two-layer MLPs. Following [19, 29], the transfer model \mathcal{T}_θ is U-Net. The transfer step $T = 1000$, the number of individualized attributes $K = 10$, the batch size $N = 64$, and the weights $\lambda_f, \lambda_r = 1$. For the n -th iteration, the epochs of two learning stages are 10 and $10n$.

4.2. Comparative Evaluation

The proposed CaPL is evaluated by comparing with 10 state-of-the-art prompt learning methods listed in Table 1. It is noted that CLIP[39] refers to directly use CLIP for evaluation without additional operation. The second to sixth methods are global prompt learning methods, while the last 4 are local methods. The results of the comparative methods are cited from their original papers. It is further noted that Tables 2-3 include some of the 10 comparative methods, as some were not evaluated on these tasks in their papers.

Base-to-new generalization. Table 1 reports the results on base-to-new generalization task, where three points can be revealed: (1) For coarse-grained datasets (e.g., [10, 43, 50]), both the global and local methods, as well as our proposed CaPL, achieve superior performance, demonstrating that the distinct differences between coarse-grained classes can be effectively captured; (2) For fine-grained datasets (e.g., [8, 27, 33, 34]), our CaPL significantly outperforms both the global and local comparative methods. For example, it improves the harmonic mean by 3.15% on Flowers102[34], 2.44% on StanfordCars[27], and 2.07% on FGVC Aircraft[33]. These results indicate the effectiveness of the proposed visual granulation technique in capturing subtle discrepancies among fine-grained classes; (3) Across the average performance of all 11 datasets, our CaPL achieves superior recognition results, further indicating the effectiveness of our method for prompt learning.

Cross-dataset transfer. As shown in Table 2, which re-

Table 1. Comparisons with the state-of-the-art CLIP-based prompt learning methods on base-to-new generalization on 11 public image recognition datasets. The best and second best results are marked in **bold** and underline.

	Average			ImageNet-1K[9]			Caltech101[10]			OxfordPets[37]		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H
CLIP[39]	69.34	74.22	71.70	72.43	68.14	70.22	96.84	94.00	95.40	91.17	97.26	94.12
CoOp[61]	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
LoGoPrompt[42]	84.47	74.24	79.03	76.74	70.83	73.66	98.19	93.78	95.93	96.07	96.31	96.18
COMMA[20]	82.42	75.87	79.04	76.04	70.89	73.86	97.94	94.56	96.50	95.62	97.84	96.72
ProMetaR[36]	84.39	76.93	80.49	77.76	70.75	74.09	98.11	94.29	96.16	95.57	97.43	96.49
TCP[55]	84.13	75.36	79.51	77.27	69.87	73.38	98.23	94.67	96.42	94.67	97.20	95.92
HPT[48]	84.32	76.86	80.23	77.95	70.74	74.17	<u>98.37</u>	94.98	96.65	95.78	97.75	96.71
CPL[58]	84.38	78.03	81.08	78.74	72.03	75.24	98.35	95.13	96.71	95.86	98.21	97.02
CoCoLe[57]	<u>85.22</u>	<u>80.31</u>	<u>82.70</u>	<u>79.25</u>	<u>74.58</u>	<u>76.84</u>	98.17	<u>95.67</u>	<u>96.90</u>	<u>96.21</u>	<u>98.55</u>	<u>97.37</u>
CDC[59]	83.34	77.38	80.25	77.50	71.73	74.51	98.20	94.37	96.25	96.07	98.00	97.02
CaPL (ours)	86.37	82.40	84.34	79.91	75.68	77.80	98.96	96.03	97.47	96.95	99.14	98.03
	StanfordCars[27]			Flowers102[34]			Food101[4]			FGVCAircraft[33]		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H
CLIP[39]	63.37	74.89	68.65	72.08	77.80	74.83	90.10	91.22	90.66	27.19	36.29	31.09
CoOp[61]	78.12	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
LoGoPrompt[42]	78.36	72.39	75.26	99.05	76.52	86.34	90.82	91.41	91.11	45.98	34.67	39.53
COMMA[20]	73.48	74.91	73.96	94.86	75.13	83.88	90.42	92.74	91.84	36.47	34.23	35.84
ProMetaR[36]	78.32	75.18	76.72	<u>98.13</u>	77.66	86.70	90.80	91.89	91.34	42.02	38.63	40.25
TCP[55]	<u>80.80</u>	74.13	77.32	97.73	75.57	85.23	90.57	91.37	90.97	41.97	34.43	37.83
HPT[48]	76.95	74.23	75.57	98.17	78.37	87.16	90.46	91.57	91.01	42.68	38.13	40.28
CPL[58]	79.31	76.65	77.96	98.07	80.43	88.38	91.92	93.87	92.88	42.27	38.85	40.49
CoCoLe[57]	80.32	<u>78.84</u>	<u>79.57</u>	97.72	<u>81.04</u>	<u>88.60</u>	<u>92.23</u>	<u>94.28</u>	<u>93.24</u>	43.86	<u>42.65</u>	<u>43.25</u>
CDC[59]	73.80	73.97	73.88	96.93	75.07	84.61	90.87	92.33	91.59	37.47	37.50	37.48
CaPL (ours)	82.92	81.11	82.01	97.44	86.68	91.75	93.25	96.87	95.03	<u>45.85</u>	44.81	45.32
	SUN397[50]			DTD[8]			EuroSAT[16]			UCF101[43]		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H
CLIP[39]	69.36	75.35	72.23	53.24	59.90	56.37	56.48	64.05	60.03	70.53	77.50	73.85
CoOp[61]	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
LoGoPrompt[42]	81.20	78.12	79.63	82.87	60.14	69.70	93.67	69.44	79.75	86.19	73.07	79.09
COMMA[20]	80.94	79.32	80.86	81.04	58.62	68.32	93.56	74.26	83.42	84.06	80.56	81.84
ProMetaR[36]	82.70	79.02	80.82	83.02	64.05	72.31	94.94	77.44	85.30	86.97	79.84	83.25
TCP[55]	82.63	78.20	80.35	72.77	58.07	68.25	91.63	74.73	82.32	87.13	80.77	83.83
HPT[48]	82.57	79.26	80.88	<u>83.84</u>	63.33	72.16	94.24	77.12	84.82	86.52	80.06	83.16
CPL[58]	81.88	79.65	80.75	80.92	62.27	70.38	94.18	81.05	87.12	86.73	80.17	83.32
CoCoLe[57]	<u>83.97</u>	<u>82.24</u>	<u>83.10</u>	82.46	<u>68.38</u>	<u>74.76</u>	95.03	<u>84.17</u>	<u>89.27</u>	<u>88.30</u>	<u>83.05</u>	<u>85.60</u>
CDC[59]	82.37	80.03	81.18	82.70	64.10	72.22	<u>95.10</u>	82.33	88.26	85.70	81.73	83.67
CaPL (ours)	84.66	83.03	83.84	84.56	71.23	77.32	95.96	85.97	90.69	89.60	85.81	87.66

ports the results on cross-dataset transfer task, our CaPL achieves the best performance on the source dataset and 9 out of 10 target datasets. Notably, it delivers significant improvements on fine-grained datasets, such as 3.16% on Flowers102[34] and 2.70% on StanfordCars[27]. These results indicate the effectiveness of our CaPL in differentiated attribute perception via the visual granulation technique.

Cross-domain generalization. Table 3 reports the results on cross-domain generalization task. As is seen, our CaPL achieves the best performance on 3 variants of ImageNet, with the second best performance on ImageNet-V2[40], demonstrating the generalizability of our proposed CaPL.

4.3. Ablation Study and Model Analysis

In this subsection, all the experiments are conducted on ImageNet-1K[9] via base-to-new generalization.

Ablation study. We conduct ablation study on the BBDM-based network (denoted as “BBDM”), factual intervention (denoted as “FAC”) and counterfactual intervention (denoted as “CON”). We also evaluate the effectiveness of in-

tegrating non-individualized attributes in factual intervention (denoted as “Non-id”). The corresponding results by omitting the four parts respectively are reported in Table 4. As is seen, all the four parts are effective in perceiving specific attributes for prompt learning, and the proposed CaPL achieves the best performance by utilizing all the four parts.

Analysis of attribute disentanglement module. First, in Table 5, we evaluate two other optimization methods besides the BBDM-based network for attribute disentanglement: (1) Classification-based method that classifies the non-individualized and individualized attribute representations by maximizing the cross entropy loss for the non-individualized representations while minimizing it for individualized representations; (2) DDPM-based method that uses the denoising diffusion model (DDPM)[19]-based network to optimize the disentangled representations. Furthermore, we evaluate a variant of our BBDM-based method (denoted as “BBDM-based variant”) that regards the individualized representations as source while the non-individualized ones as condition. Four points can be re-

Table 2. Comparisons with comparative prompt learning methods on cross-dataset transfer. ImageNet-1K is used as source and the others are target. The datasets are denoted by abbreviations for simplicity. The best and second best results are marked in **bold** and underline.

	Source					Target							
	ImageNet.[9]	Caltech.[10]	Pets.[37]	Cars.[27]	Flowers.[34]	Food.[4]	Aircraft.[33]	SUN.[50]	DTD[8]	Euro.[16]	UCF.[43]	Avg.	
CoOp[61]	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88	
COMMA[20]	71.22	93.84	90.78	66.36	73.14	85.87	25.14	67.56	46.52	48.85	68.71	66.84	
TCP[55]	71.40	93.97	91.25	64.69	71.21	86.69	23.45	67.15	44.35	51.45	68.73	66.29	
HPT[48]	71.72	94.20	<u>92.63</u>	66.33	<u>74.84</u>	86.21	25.68	<u>68.75</u>	<u>50.87</u>	47.36	70.50	67.74	
CPL[58]	73.53	95.52	91.64	66.17	73.35	87.68	27.36	68.24	48.96	51.25	70.52	68.07	
CoCoLe[57]	<u>73.88</u>	95.88	91.93	<u>67.79</u>	74.17	<u>87.97</u>	<u>28.83</u>	<u>68.75</u>	49.26	<u>51.75</u>	<u>72.78</u>	<u>68.91</u>	
CDC[59]	71.76	94.47	90.77	66.27	72.67	86.27	24.50	68.07	46.60	49.13	68.60	66.73	
CaPL (ours)	75.02	<u>95.78</u>	92.81	70.49	78.00	89.02	30.88	69.39	51.22	52.95	73.91	70.45	

Table 3. Comparisons with comparative prompt learning methods on cross-domain generalization. ImageNet-1K is used as source dataset and its 4 variants are used as target datasets. The best and second best results are marked in **bold** and underline.

	Source	Target				
	ImageNet.[9]	-V2[40]	-S[46]	-A[18]	-R[17]	Avg.
CoOp[61]	71.51	64.20	47.99	49.71	75.21	59.28
LoGoPrompt[42]	75.27	66.65	48.99	51.36	76.85	60.96
COMMA[20]	71.22	64.84	49.65	51.64	77.56	60.92
ProMetaR[36]	71.29	64.39	49.55	51.25	77.89	60.77
HPT[48]	71.72	65.25	49.36	50.85	77.38	60.71
CPL[58]	73.53	65.18	49.92	50.73	77.38	60.80
CoCoLe[57]	73.88	65.86	<u>50.89</u>	<u>51.75</u>	<u>78.89</u>	<u>61.85</u>
CDC[59]	71.76	64.87	50.33	50.40	78.10	60.93
CaPL (ours)	<u>75.02</u>	<u>66.63</u>	51.47	53.32	79.38	62.70

Table 4. Ablation study for different parts of CaPL.

Method	Base	New	H
CaPL w/o BBDM	77.90	73.94	75.87
CaPL w/o UNI	77.32	73.58	75.40
CaPL w/o CON	78.79	72.32	75.42
CaPL w/o Non-id	78.58	74.59	76.53
CaPL (ours)	79.91	75.68	77.80

vealed: (1) The classification-based method outperforms the method without additional optimization (*i.e.*, CaPL w/o BBDM in Table 4), showing the effectiveness of enforcing constraints on the disentangled representations; (2) The DDPM-based method performs worse than the two BBDM-based methods, since DDPM models the non-individualized representations as a noise distribution, which negatively impacts disentanglement accuracy; (3) The BBDM-based variant underperforms ours, mainly because the shared non-individualized representations are better suited as a common distribution for all visual features, while the class-specific individualized ones serve more effectively as conditions for guiding feature transitions to specific visual features; (4) The proposed BBDM-based network achieves the best performance, indicating the effectiveness of our proposed method to promote attribute disentanglement.

Then, we visualize the disentangled attribute representations in Fig. 5(a), where the non-individualized representations (denoted as “o”) and individualized representations (denoted as “□”) disentangled from 10 images from the fine-grained StanfordCars dataset[27] are visualized, where each image belongs to a different class. As is seen,

Table 5. Ablation study for different optimization methods for attribute disentanglement.

Method	Base	New	H
Classification-based method	78.44	74.24	76.28
DDPM-based method	79.12	74.93	76.97
BBDM-based variant	79.33	74.89	77.05
BBDM-based method (ours)	79.91	75.68	77.80

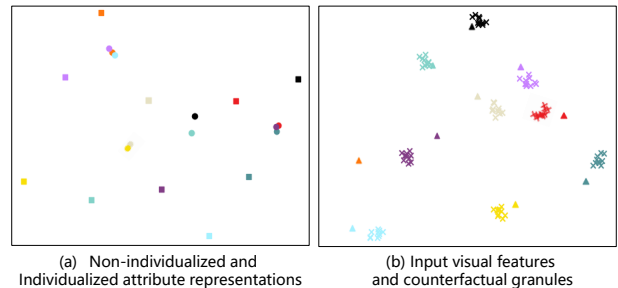


Figure 5. T-SNE visualizations of 10 images from 10 different classes of the StanfordCars dataset[27]. (a) visualizes the non-individualized attribute representations (denoted as “o”) and individualized attribute representations (denoted as “□”). (b) visualizes the input visual features (denoted as “Δ”) of the 10 images, and the counterfactual granules (denoted as “x”) constructed by swapping the attributes across the 10 images.

the non-individualized representations are grouped into several small clusters, illustrating that these attributes have the same value across some classes but differ in other classes. In contrast, the individualized representations are scattered, further demonstrating that the attributes with different discrimination ability are well disentangled.

Analysis of factual intervention. First, we present visualizations to evaluate the effectiveness of the attribute queries in extracting representations of single individualized attributes. An image sample from the StanfordCars dataset[27] is shown in Fig. 6(a), and its corresponding visual representations of all individualized attributes are visualized in Fig. 6(b) by utilizing the visualization method in [63] (Note: The number of individualized attributes is set to 10 in our experiments). As is seen, each attribute query could extract an individualized attribute-specific visual representation, focusing on a discriminative part of the



Figure 6. An image sample (a) and the attention maps (b) of its corresponding individualized attribute representations.

Table 6. Ablation study for attribute query initialization.

Method	Base	New	H
Non-learnable method	77.83	72.75	75.20
Prior initialization method	78.34	74.12	76.17
Random initialization method (ours)	79.91	75.68	77.80

Table 7. Ablation study for three hyperparameters.

Value of K	0	2	4	6	8	10	12
Accuracy (H)	75.66	75.84	76.39	76.78	77.23	77.80	77.64
Value of λ_f	0	0.2	0.4	0.6	0.8	1	1.2
Accuracy (H)	76.99	77.52	77.61	77.71	77.77	77.80	77.78
Value of λ_r	0	0.2	0.4	0.6	0.8	1	1.2
Accuracy (H)	75.41	76.02	76.83	77.29	77.60	77.80	77.39

target object. Furthermore, it is noted that the last visual representation does not contain clear information, likely because the corresponding individualized attribute relates to the car’s tail, which is not visible in the image. This suggests that when an image lacks relevant information, the attribute query can extract a representation indicating the absence of clear information, further demonstrating its ability to extract attribute-specific representations.

Then, considering that the attribute queries are randomly initialized, we evaluate two other initialization methods in Table 6: (1) Non-learnable method that uses GPT-3[5] to generate 10 phrases describing 10 individualized attributes that are discriminative for recognizing a class, and encodes these phrases by CLIP text encoder to act as fixed queries; (2) Prior initialization method that initializes the learnable attribute queries with the above queries obtained by GPT-3. The results in Table 6 demonstrate the effectiveness of the proposed random initialization method in learning accurate attribute queries. It is noted that the prior initialization method underperforms the random method, mainly due to that random initialization could provide better coverage, improving representation and generalization in high-dimensional spaces as indicated in [15].

Analysis of counterfactual intervention. We evaluate the effectiveness of constructing counterfactual granules in counterfactual intervention. Specifically, we select 10 images from 10 classes of StanfordCars[27], and construct the corresponding 100 counterfactual granules. Each granule is assigned the same class as the image from which its individualized attribute is disentangled. In Fig. 5(b), we visualize the input visual features (denoted as “ Δ ”) and counterfactual granules (denoted as “ \times ”), with different colors indicating different classes. As shown, counterfactual gran-

ules that share the same individualized attribute (*i.e.*, same color) cluster around the corresponding visual features from which the individualized attribute is extracted, accurately reflecting the grouping of granules with the same class label near their real counterparts. This demonstrates the effectiveness of (1) the disentanglement of non-individualized and individualized attributes, (2) the decoder in generating counterfactual granules, and (3) the simulation of alternative contexts for improving generalization.

Hyperparameters. We analyze the number of individualized attributes K and the weights λ_f, λ_r in Table 7. Three points can be revealed: (1) Increasing K improves accuracy, as disentangling more individualized attributes enhances finer perception. The performance stabilizes when $K > 10$, indicating sufficient individualized attributes for recognition, so we set $K = 10$; (2) The accuracy is insensitive to λ_f , so we set $\lambda_r = 1$; (3) The best performance is achieved at $\lambda_r = 1$, as it balances reconstruction error and cross entropy loss, so we set $\lambda_f = 1$.

Limitation. Our CaPL has a relatively long training time due to the use of BBDM-based attribute disentanglement module. However, our method only uses the text prompt to calculate cosine similarity for inference, resulting in a faster inference time than some comparative methods while matching that of the others.

5. Conclusion

In this paper, we construct an attribute-driven prompt learning graph for recognition, which depicts the relationship between a text prompt, visual features and attributes with different discrimination ability from a causal perspective. Accordingly, we propose a causality-guided text prompt learning method, where a visual granulation technique could capture subtle discrepancies among fine-grained classes by constructing sets of visual granules as supervision by integrating attributes under two causal inference strategies. Experimental results have demonstrated the superiority of the proposed method. In future, we would investigate how to construct more accurate granules for prompt learning.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (Grant Nos. 62376269, 61991423, U1805264), the Beijing Municipal Science and Technology Project (Grant No. Z211100011021004).

References

- [1] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10044–10054, 2020. 4
- [2] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5
- [3] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 2
- [4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *Computer Vision—ECCV 2014: 13th European Conference*, pages 446–461, 2014. 6, 7, 1, 2
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. 8
- [6] Ziliang Chen, Xin Huang, Quanlong Guan, Liang Lin, and Weiqi Luo. A retrospect to multi-prompt learning across vision and language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22190–22201, 2023. 2
- [7] Youngjae Cho, HeeSun Bae, Seungjae Shin, Yeo Dong Youn, Weonyoung Joo, and Il-Chul Moon. Make prompts adaptable: Bayesian modeling for vision-language prompt learning with data-dependent prior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11552–11560, 2024. 2
- [8] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 5, 6, 7, 1, 2
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5, 6, 7, 1, 2
- [10] 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 *2004 Conference on Computer Vision and Pattern Recognition workshop*, pages 178–178. IEEE, 2004. 5, 6, 7, 1, 2
- [11] Jiannan Ge, Hongtao Xie, Shaobo Min, Pandeng Li, and Yongdong Zhang. Dual part discovery network for zero-shot learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3244–3252, 2022. 1, 2
- [12] Yunhao Ge, Jie Ren, Andrew Gallagher, Yuxiao Wang, Ming-Hsuan Yang, Hartwig Adam, Laurent Itti, Balaji Lakshminarayanan, and Jiaping Zhao. Improving zero-shot generalization and robustness of multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11093–11101, 2023. 1
- [13] Yiwei Guo, Shaobin Zhuang, Kunchang Li, Yu Qiao, and Yali Wang. Transagent: Transfer vision-language foundation models with heterogeneous agent collaboration. *Advances in Neural Information Processing Systems*, 2025. 1, 2
- [14] Tianxiang Hao, Xiaohan Ding, Juexiao Feng, Yuhong Yang, Hui Chen, and Guiguang Ding. Quantized prompt for efficient generalization of vision-language models. In *European Conference on Computer Vision*, pages 54–73, 2025. 1, 2
- [15] Kun He, Yan Wang, and John Hopcroft. A powerful generative model using random weights for the deep image representation. *Advances in Neural Information Processing Systems*, 29, 2016. 8
- [16] 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. 6, 7, 1, 2
- [17] 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/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 7, 1, 2
- [18] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 7, 1, 2
- [19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 5, 6
- [20] Lianyu Hu, Liqing Gao, Zekang Liu, Chi-Man Pun, and Wei Feng. Comma: Co-articulated multi-modal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2238–2246, 2024. 1, 2, 6, 7
- [21] Chen Huang, Skyler Seto, Samira Abnar, David Grangier, Navdeep Jaitly, and Josh Susskind. Aggregate-and-adapt natural language prompts for downstream generalization of clip. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 1, 2
- [22] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727, 2022. 2
- [23] Baoshuo Kan, Teng Wang, Wenpeng Lu, Xiantong Zhen, Weili Guan, and Feng Zheng. Knowledge-aware prompt tuning for generalizable vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15670–15680, 2023. 2
- [24] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 2
- [25] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational

- model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023. 2
- [26] Gahyeon Kim, Sohee Kim, and Seokju Lee. Aapl: Adding attributes to prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1572–1582, 2024. 1
- [27] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 5, 6, 7, 8, 1, 2, 3
- [28] Marc Lafon, Elias Ramzi, Clément Rambour, Nicolas Audebert, and Nicolas Thome. Gallop: Learning global and local prompts for vision-language models. In *European Conference on Computer Vision*, pages 264–282, 2024. 1, 2
- [29] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdt: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1952–1961, 2023. 2, 3, 5
- [30] Tang Li, Mengmeng Ma, and Xi Peng. Deal: Disentangle and localize concept-level explanations for vlms. In *European Conference on Computer Vision*, pages 383–401, 2024. 1, 2
- [31] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26617–26626, 2024. 1
- [32] Zonglin Lyu, Ming Li, Jianbo Jiao, and Chen Chen. Frame interpolation with consecutive brownian bridge diffusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3449–3458, 2024. 3
- [33] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1, 5, 6, 7, 2
- [34] 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, 2008. 1, 5, 6, 7, 2
- [35] Yassine Ouali, Adrian Bulat, Brais Matinez, and Georgios Tzimiropoulos. Black box few-shot adaptation for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15534–15546, 2023. 2
- [36] Jinyoung Park, Juyeon Ko, and Hyunwoo J Kim. Prompt learning via meta-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26940–26950, 2024. 1, 2, 5, 6, 7
- [37] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505. IEEE, 2012. 6, 7, 1, 2, 3
- [38] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023. 1
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 1, 2, 5, 6
- [40] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400, 2019. 6, 7, 1, 2
- [41] Karsten Roth, Jae Myung Kim, A Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15746–15757, 2023. 1
- [42] Cheng Shi and Sibeil Yang. Logoprompt: Synthetic text images can be good visual prompts for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2932–2941, 2023. 1, 2, 6, 7
- [43] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5, 6, 7, 1, 2
- [44] Simon Vandenhende, Dhruv Mahajan, Filip Radenovic, and Deepti Ghadiyaram. Making heads or tails: Towards semantically consistent visual counterfactuals. In *European Conference on Computer Vision*, pages 261–279, 2022. 4
- [45] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 5
- [46] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 7, 1, 2
- [47] Peiyong Wang, Bohan Xiao, Qisheng He, Carri Glide-Hurst, and Ming Dong. Score-based image-to-image brownian bridge. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10765–10773, 2024. 3
- [48] Yubin Wang, Xinyang Jiang, De Cheng, Dongsheng Li, and Cairong Zhao. Learning hierarchical prompt with structured linguistic knowledge for vision-language models. In *Proceedings of the AAI Conference on Artificial Intelligence*, pages 5749–5757, 2024. 1, 2, 6, 7
- [49] Zhengbo Wang, Jian Liang, Ran He, Nan Xu, Zilei Wang, and Tieniu Tan. Improving zero-shot generalization for clip with synthesized prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3032–3042, 2023. 2
- [50] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010. 5, 6, 7, 1, 2
- [51] Chen Xu, Yuhan Zhu, Haocheng Shen, Boheng Chen, Yixuan Liao, Xiaoxin Chen, and Limin Wang. Progressive visual prompt learning with contrastive feature re-formation.

International Journal of Computer Vision, pages 1–16, 2024.
[2](#)

- [52] Shiyu Xuan, Ming Yang, and Shiliang Zhang. Adapting vision-language models via learning to inject knowledge. *IEEE Transactions on Image Processing*, 2024. [1](#), [2](#)
- [53] Tao Yang, Cuiling Lan, Yan Lu, and Nanning Zheng. Diffusion model with cross attention as an inductive bias for disentanglement. *Advances in Neural Information Processing Systems*, 2025. [3](#)
- [54] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023. [1](#), [2](#)
- [55] Hantao Yao, Rui Zhang, and Changsheng Xu. Tcpl: Textual-based class-aware prompt tuning for visual-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23438–23448, 2024. [1](#), [2](#), [5](#), [6](#), [7](#)
- [56] Sihui Zhang and Zhijiang Li. Class-aware visual prompt learning for vision language models. *Journal of Imaging Science and Technology*, pages 1–7, 2025. [2](#)
- [57] Yi Zhang, Ke Yu, Siqi Wu, and Zhihai He. Conceptual codebook learning for vision-language models. In *European Conference on Computer Vision*, pages 235–251, 2024. [1](#), [2](#), [5](#), [6](#), [7](#)
- [58] Yi Zhang, Ce Zhang, Ke Yu, Yushun Tang, and Zhihai He. Concept-guided prompt learning for generalization in vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7377–7386, 2024. [1](#), [2](#), [6](#), [7](#)
- [59] Yanan Zhang, Jiangmeng Li, Lixiang Liu, and Wenwen Qiang. Rethinking misalignment in vision-language model adaptation from a causal perspective. *Advances in Neural Information Processing Systems*, 2025. [1](#), [2](#), [6](#), [7](#)
- [60] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. [5](#)
- [61] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. [1](#), [2](#), [4](#), [6](#), [7](#)
- [62] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15659–15669, 2023. [1](#), [2](#)
- [63] Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2605–2615, 2023. [1](#), [2](#), [7](#)
- [64] Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, and Hanwang Zhang. Enhancing zero-shot vision models by label-free prompt distribution learning and bias correcting. *Advances in Neural Information Processing Systems*, 37:2001–2025, 2024. [1](#)