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

Supplementary Material

A. Dataset Details

In this paper, we use 11 public recognition datasets for the base-to-new generalization task and cross-dataset transfer task, including ImageNet-1K[9], Caltech101[10], OxfordPets[37], StanfordCars[27], Flowers102[34], Food101[4], FGVCAircraft[33], SUN397[50], DTD[8], EuroSAT[16], and UCF101[43]. The detailed statistics of these datasets are listed in Table S1.

Furthermore, we use 4 variants of ImageNet[9] for the cross-domain generalization task, including ImageNet-V2[40], ImageNet-S[46], ImageNet-A[18], and ImageNet-R[17]. Each variant contains the same classes to ImageNet[9] but different image distributions. The detailed statistics of these datasets are listed in Table S2. It is noted that ImageNet-V2[40] and ImageNet-S[46] contain all the 1,000 classes of ImageNet, while ImageNet-A[18] and ImageNet-R[17] select 200 classes from the 1000 classes of ImageNet.

B. More Visualization Results

First, in Figs. S1-S2, we provide two more image samples from the fine-grained StanfordCars dataset[27] and the fine-grained OxfordPets dataset[37] respectively to visualize the attention maps of the visual representations of single individualized attributes extracted by the attribute queries.

As seen from Fig. S1, the visual representations extracted by the attribute queries focus on similar attributes to that shown in Fig. 6 in our main paper, and the visual representations (*e.g.*, the seventh to ninth attention maps) do not contain clear information when the image lacks relevant information.

Then, as seen from Fig. S2, the attribute queries could extract visual representations paying attention to the discriminative attributes for recognizing a pet, such as eyes, ears, and tail. It should be noted that the seventh and eighth attention maps focus on the same region while containing different information. This is mainly because that the visualization of attention maps can only provide an intuitive understanding of the regions associated with specific attributes but cannot explicitly determine which attribute is being represented. Therefore, the seventh attention map likely captures the texture of the dog's fur, as the image contains relevant texture information, making this attention map more informative. In contrast, the eighth attention map does not convey clear information, suggesting that it may correspond to an attribute such as the pattern on the dog's body—an attribute that is absent in this particular image. The above observations are consistent with that observed in Fig. 6, which further demonstrate the effectiveness of the attribute queries in extract individualized attribute-specific representations.

Furthermore, in Fig. S3, we present heatmap visualizations to evaluate the effectiveness of the disentangled attributes and the counterfactual granules by utilizing 10 images from 10 different classes of the StanfordCars dataset[27]. Specifically, with the non-individualized and individualized attribute representations disentangled from the 10 image visual features, we calculate the cosine similarity within the 10 non-individualized attribute representations and the 10 individualized attribute representations respectively. The corresponding heatmaps of these cosine similarities are visualized in Figs. S3(a)-(b) respectively. As seen from Fig. S3(a), non-individualized attribute representations exhibit similarity among some classes while remaining different from others. For example, the non-individualized representation of the first class is nearly identical to that of the ninth class and also shares similarities with the third to fifth and seventh to eighth classes. However, it differs from the representations of the second and sixth classes. This demonstrates that the non-individualized attributes carry weak discrimination ability. As seen from Fig. S3(b), the individualized attribute representations of different classes are distinct to each other, demonstrating that the individualized attributes have strong discrimination ability for distinguishing one class from the other classes. The observations from Figs. S3(a)-(b) are consistent to Fig. 5(a) in our main paper, further demonstrating the effectiveness of the attribute disentanglement according to discrimination ability.

To further evaluate the effectiveness of the counterfactual granules, we randomly select an individualized attribute representation from the 10 images, and integrate it with all the 10 non-individualized attribute representations to construct 10 counterfactual granules. The cosine similarity within these counterfactual granules are calculated, and the heatmap of these cosine similarities is visualized in Fig. S3(c). Similarly, we calculate and visualize the cosine similarity within the other 10 counterfactual granules that share the non-individualized attribute but have different individualized attributes in Fig. S3(d). As seen from these two figures, the counterfactual granules with different non-individualized attributes are similar across some classes, while the counterfactual granules with different individualized attributes are distinct to each other, which are

Table S1. Statistics of the 11 public recognition datasets.

Dataset	Publication information	Number of classes	Number of images for training	Number of images for testing	Task
ImageNet-1K[9]	CVPR 2019	1000	1,281,167	50,000	General object recognition
Caltech101[10]	CVPRW 2004	101	4,128	2,465	General object recognition
OxfordPets[37]	CVPR 2012	37	2,944	3,669	Fine-grained pet recognition
StanfordCars[27]	ICCVW 2013	196	6,509	8,041	Fine-grained car recognition
Flowers102[34]	ICVGIP 2008	102	4,093	2,463	Fine-grained flower recognition
Food101[4]	ECCV 2014	101	50,500	30,300	Fine-grained food recognition
FGVCAircraft[33]	arXiv 2013	100	3,334	3,333	Fine-grained aircraft recognition
SUN397[50]	CVPR 2010	397	15,880	19,850	Scene recognition
DTD[8]	CVPR 2014	47	2,820	1,692	Texture recognition
EuroSAT[16]	JSTARS 2019	10	13,500	8,100	Satellite image recognition
UCF101[43]	arXiv 2012	101	7,639	3,783	Action recognition

Table S2. Statistics of the 4 variants of ImageNet[9].

Dataset	Publication information	Image distribution	Number of images
ImageNet-V2[40]	ICML 2019	New images following the same distribution as ImageNet[9]	30,000
ImageNet-S[46]	NeurIPS 2019	Black and white sketches	50,000
ImageNet-A[18]	CVPR 2021	Adversarial images with subtle disturbance	7,500
ImageNet-R[17]	ICCV 2021	Artificial images	30,000

Table S3. Comparison of the training and inference times with comparative methods. The training and inference of the comparative methods and our proposed method are conducted on the StanfordCars dataset[27] by utilizing one NVIDIA RTX A5000. "h" denotes one hour, and "s" denotes one second.

Method	Training time	Inference time	H (harmonic mean)
CoOp[61]	1.86h	0.24s	68.13
LoGoPrompt[42]	3.12h	0.24s	75.26
COMMA[20]	2.34h	0.47s	73.96
TCP[55]	2.48h	0.34s	77.32
CPL[58]	4.21h	0.24s	77.96
CoCoLe[57]	4.21h	2.90s	79.57
CaPL (ours)	4.51h	0.24s	82.01

consistent with Fig. 5(b) in our main paper, further demonstrating the effectiveness of the disentangled attributes and the simulation of alternative context to alleviate spurious correlations.

C. Efficiency Analysis

At the training stage, the proposed method adopts a two-stage training scheme, which involves learning a BBDM for attribute disentanglement, leading to a relatively long training time. In contrast, at the inference stage, our method only uses the learned text prompt to calculate cosine similarity for recognition, resulting in a short inference time. Specifically, Table S3 presents the training and inference times of several comparative methods and our proposed CaPL on StanfordCars[27] under the base-to-new generalization setting. The results for the comparative methods are obtained using their released codes and official implementation details, and all experiments are conducted on one NVIDIA RTX A5000. Additionally, we report the harmonic mean of each method. Three key observations can be revealed from the table: (1) The training times of the comparative methods [20, 55, 61] are relatively short, since they primarily learn prompts (and, in some cases, simple additional modules); (2) The training times of the comparative methods [42, 57, 58] and our proposed CaPL are relatively long due to the two-stage training scheme, which involves learning additional components such as a name-to-image generator [42], a visual cache [58], a conceptual codebook [57], and a BBDM (our CaPL), alongside prompt learning; (3) The inference time of our CaPL is as short as that of [42, 58, 61], since our CaPL and these comparative methods only involve calculating cosine similarity using the learned prompt. In contrast, methods like [20, 55, 57] incur longer inference times due to additional modules and operations. The combination of comparable training time, short inference time, and the highest harmonic mean highlights the efficiency and effectiveness of our proposed CaPL.

(a) Images (b) Attention maps of visual representations of single individualized attributes

Figure S1. An image sample (a) from the StanfordCars dataset[27] and the attention maps (b) of its corresponding individualized attribute representations.

(a) Images (b) Attention maps of visual representations of single individualized attributes

Figure S2. An image sample (a) from the OxfordPets dataset[37] and the attention maps (b) of its corresponding individualized attribute representations.

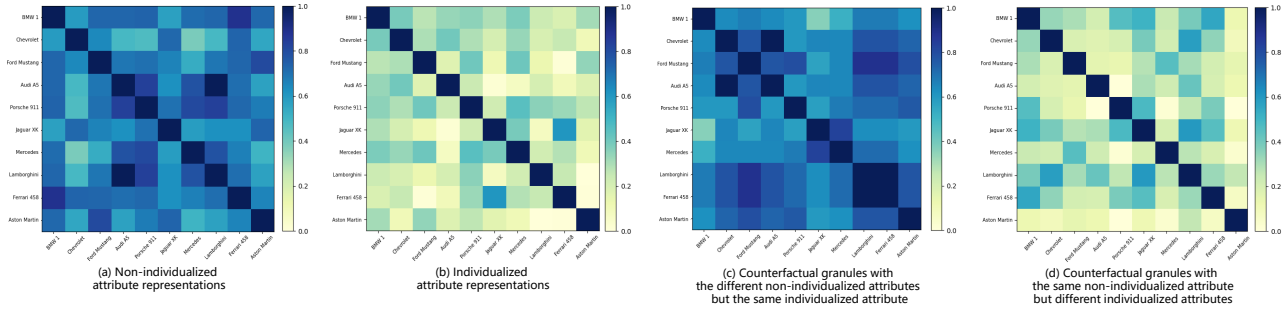

Figure S3. Heatmaps of the cosine similarities calculated within (a) non-individualized attribute representations, (b) individualized attribute representations, (c) counterfactual granules that have different non-individualized attributes but share the individualized attribute, and (d) counterfactual granules that share the non-individualized attribute but have different individualized attributes.