A Conditional Probability Framework for Compositional Zero-shot Learning Supplementary Materials

Peng Wu[1]*, Qiuxia Lai[2]*, Hao Fang[1], Guo-Sen Xie[3], Yilong Yin[1], Xiankai Lu[1]†, Wenguan Wang[4,5]

[1]Shandong University, [2] Communication University of China, [3]Nanjing University of Science and Technology,

[4]Zhejiang University, [5]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University

https://github.com/Pieux0/CPF

This document offers theoretical justification and implementation of CPF, pseudo-code of CPF, additional experimental results, further qualitative analysis, comprehensive implementation details and extensive dataset information. The structure is organized as follows:

1. Theoretical Justification and Implementation of CPF

The chain rule of probability universally decomposes any joint probability into a product of conditional and marginal probabilities [3]: $p(a, o|x) = p(a|o, x)p(o|x)$ (or symmetrically $p(a, o|x) = p(o|a, x)p(a|x)$). In CZSL, the choice to decompose $p(a, o|x)$ as $p(a|o, x)p(o|x)$ (rather than $p(a|x)p(o|x)$ or $p(o|a, x)p(a|x)$) is driven by semantic and contextual dependencies: (i) the independence assumption $p(a)p(o)$ fails to capture semantic binding [4, 10], whereas the conditional probability explicitly models plausible attribute-object relationships. (ii) The semantic asymmetry (established empirically by Nagarajan and Grauman [9]) structurally favors $p(a|o, x)$ over $p(o|a, x)$, as objects act as semantic anchors that causally determine plausible attributes, whereas the inverse mapping $p(o|a, x)$ is ill-posed due to attribute-sharing across objects, violating injectivity for stable inference. In practical, we adopt the additive formulation to address the practical issue of "probability vanishing" inherent in multiplicative approaches.

2. Pseudo-code of CPF

Algorithm 1 provides the pseudo-code of CPF.

*Equal Contribution.

†Corresponding author: *Xiankai Lu*.

Algorithm 1 Pseudo-code of CPF in a PyTorch-like style.

```
"""
# v_h_c: deep-level class feature (1 x D)
# V_h_p: deep-level patch feature (HW x D)
# v_l_c: shallow-level class feature (1 x D)
# V_h_p: shallow-level patch feature (HW X D)
# W_a: attribute textual embedding (M x D)
# W_o: object textual embedding (N X D)
# W_f_o: projection matrix (D x d)
# W_f_a: projection matrix (D x d)
# D: visual feature dimension
# d: textual embedding dimension
"""

#======= text-enhaced object learning =======#
# projecting deep-level class feature into the
    joint semantic space
v_h_c_d = v_h_c @ W_f_o
# compute similarity score
score_1 = torch.matmul(v_h_c_d, W_o.transpose()) /
     torch.sqrt( torch.tensor(d))
# compute q_t with similarity score
q_t = F.softmax(score_1, dim=-1) @ W_o
# projecting deep-level patch feature into the
    joint semantic space
V_h_p_d = V_h_p @ W_f_o
# compute attention score
score_2 = torch.matmul(q_t, V_h_p.transpose()) /
    torch.sqrt( torch.tensor(d))
# compute object feature
v_o = v_h_c + F.softmax(score_2, dim=-1) @ V_h_p

#======= object-guided attribute learning =======#
# compute attention score
score_3 = torch.matmul(v_o, V_l_p.transpose()) /
    torch.sqrt( torch.tensor(D))
# compute attribute feature
v_a = F.softmax(score_3, dim=-1) @ V_l_p
```

3. More Experiment Results

In this section, we present the ablation study on loss weight coefficients α_1 and α_2 on UT-Zappos50K [12]. The results are shown in Table 1. Additionally, we conduct ablation experiments on block choices on UT-Zappos50K [12] to select the most suitable blocks as shallow-level visual embeddings. The corresponding results are presented in Table 2. Moreover, as shown in the Table 3, our CLIP-based CPF consistently outperforms other methods on MiT-States and UT-Zappos50K.

Table 1. Ablation study on loss weight coefficients α_1 and α_2 on UT-Zappos50K [12].

α_1	α_2	UT-Zappos50K AUC↑	HM↑	Seen↑	Unseen↑
0.0	0.0	28.5	45.1	56.4	60.9
0.3	0.7	39.2	53.3	65.3	71.3
0.4	0.6	39.3	53.4	64.3	**72.5**
0.5	0.5	40.1	55.3	65.6	69.4
0.6	0.4	**41.4**	**55.7**	**66.4**	71.1
0.7	0.3	40.0	54.5	66.0	69.3

Table 2. Ablation study on block choices on UT-Zappos50K [12].

Setting	Blocks	UT-Zappos50K AUC↑	HM↑	Seen↑	Unseen↑
Close World	(1,4,7)	38.3	52.4	65.5	71.1
	(2,5,8)	39.2	53.6	65.3	**71.7**
	(3,6,9)	**41.4**	**55.7**	**66.4**	71.1
Open World	(1,4,7)	28.8	45.6	64.1	52.3
	(2,5,8)	29.0	46.2	**65.3**	51.0
	(3,6,9)	**31.2**	**47.6**	64.6	**56.1**

Table 3. Results of CLIP-based CPF on MiT-States [6] and UT-Zappos50KK [12].

Method	UT-Zappos50K AUC↑	HM↑	MiT-States AUC↑	HM↑
CDS-CZSL [7]	39.5	52.7	22.4	39.2
Troika [5]	41.7	54.6	22.1	39.3
PLID [1]	38.7	52.4	22.1	39.0
CAILA [13]	44.1	57.0	23.4	39.9
Ours	**45.2**	**57.6**	23.2	**40.5**

4. More Qualitative Analysis

We provide more qualitative results of UT-Zappos50K [12], MIT-States [6] and C-GQA [8] under *CW* and *OW* settings in Fig. 5. We show results for each dataset in each row. Images predicted under the *CW* setting are shown in the first three columns and the rest of the columns show the instances under the *OW* setting. In Fig. 1, we provide attention visualization for Eq. 2 and Eq. 4. In Fig. 2, we illustrate qualitative results of image retrieval. In Fig. 3, we show additional wrong predictions of instances in C-GQA [8].

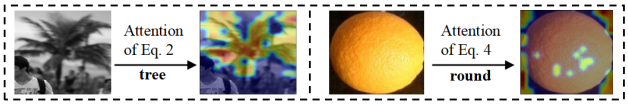

Figure 1. Attention visualizations for Eq. 2 and Eq. 4

Figure 2. Qualitative results of image retrieval

Figure 3. More qualitative results of wrong predictions of instances in C-GQA [8].

5. More Implementations Details

We provide the implementation details of deep-level and shallow-level feature extraction in Fig. 4.

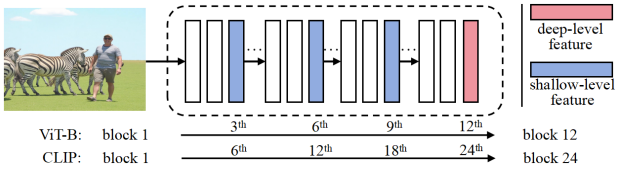

Figure 4. Implementation details of deep-level and shallow-level visual embeddings for both ViT-B and CLIP.

6. Summary of Data Split Statistics

Following previous work [2, 11], we provide the summary of data split statistics for UT-Zappos50K [12], MIT-States [6] and C-GQA [8] in Table 4. $|\mathcal{A}|$ and $|\mathcal{O}|$ represent the numbers of attribute and object classes, respectively. $|\mathcal{C}_s|$ and

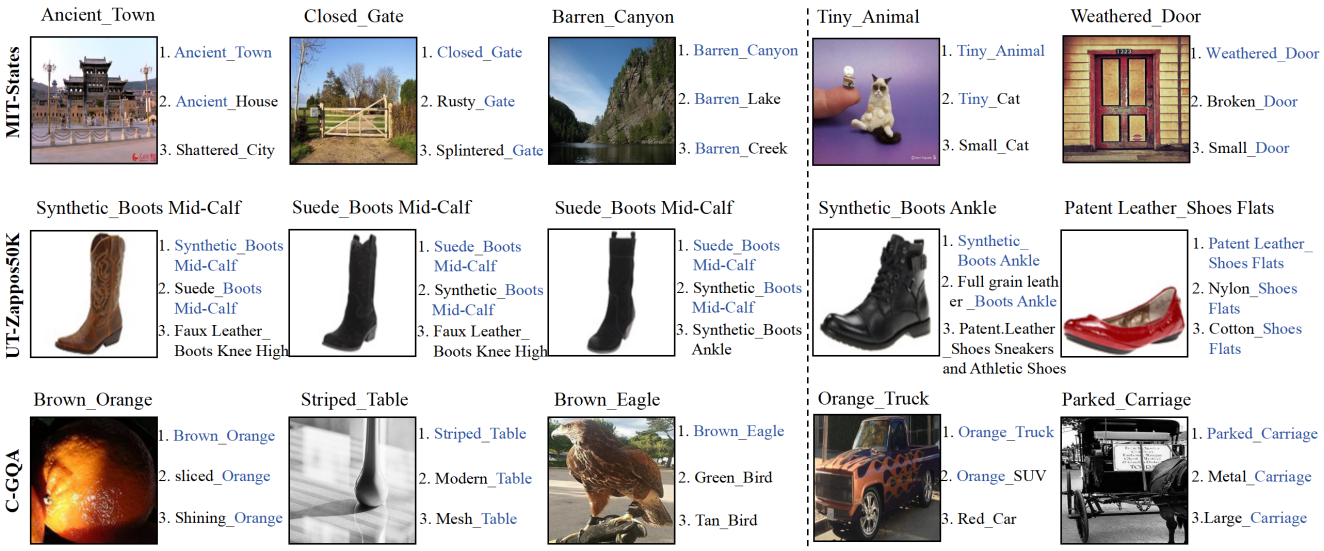

Figure 5. More qualitative results of UT-Zappos50K [12], MIT-States [6] and C-GQA [8].

Table 4. Summary of data split statistics.

	Composition			Train		Val		Test	
Datasets	$\lvert\mathcal{A}\rvert$	$\lvert\mathcal{O}\rvert$	$\lvert\mathcal{A}\rvert \times \lvert\mathcal{O}\rvert$	$\lvert\mathcal{C}_s\rvert$	$\lvert\mathcal{X}\rvert$	$\lvert\mathcal{C}_s\rvert$ / $\lvert\mathcal{C}_u\rvert$	$\lvert\mathcal{X}\rvert$	$\lvert\mathcal{C}_s\rvert$ / $\lvert\mathcal{C}_u\rvert$	$\lvert\mathcal{X}\rvert$
UT-Zappos50K	16	12	192	83	22998	15 / 15	3214	18 / 18	2914
MIT-States	115	245	28175	1262	30338	300 / 300	10420	400 / 400	12995
C-GQA	413	674	278362	5592	26920	1252 / 1040	7280	888 / 923	5098

$\lvert\mathcal{C}_u\rvert$ denote the numbers of seen and unseen composition categories, respectively. $\lvert\mathcal{X}\rvert$ indicates the numbers of images.

References

[1] Wentao Bao, Lichang Chen, Heng Huang, and Yu Kong. Prompting language-informed distribution for compositional zero-shot learning. In *ECCV*, 2024. 2

[2] Shaozhe Hao, Kai Han, and Kwan-Yee K Wong. Learning attention as disentangler for compositional zero-shot learning. In *CVPR*, 2023. 2

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 1

[4] Taihang Hu, Linxuan Li, Joost van de Weijer, Hongcheng Gao, Fahad Shahbaz Khan, Jian Yang, Ming-Ming Cheng, Kai Wang, and Yaxing Wang. Token merging for training-free semantic binding in text-to-image synthesis. *NeurIPS*, 2024. 1

[5] Siteng Huang, Biao Gong, Yutong Feng, Min Zhang, Yiliang Lv, and Donglin Wang. Troika: Multi-path cross-modal traction for compositional zero-shot learning. In *CVPR*, 2024. 2

[6] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015. 2, 3

[7] Yun Li, Zhe Liu, Hang Chen, and Lina Yao. Context-based and diversity-driven specificity in compositional zero-shot learning. *CVPR*, 2024. 2

[8] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *CVPR*, 2021. 2, 3

[9] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *ECCV*, 2018. 1

[10] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *NeurIPS*, 2023. 1

[11] Qingsheng Wang, Lingqiao Liu, Chenchen Jing, Hao Chen, Guoqiang Liang, Peng Wang, and Chunhua Shen. Learning conditional attributes for compositional zero-shot learning. In *CVPR*, 2023. 2

[12] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014. 1, 2, 3

[13] Zhaoheng Zheng, Haidong Zhu, and Ram Nevatia. Caila: Concept-aware intra-layer adapters for compositional zero-shot learning. In *WACV*, 2024. 2