Learning Visual Proxy for Compositional Zero-Shot Learning

Supplementary Material

1. Experiments Setting

Table S1. The statistics of four CZSL datasets.

Dataset			Train		Val			Test		
	$\mathcal{A}$	$\mathcal{O}$	$\mathcal{Y}_s$	$\mathcal{X}$	$\mathcal{Y}_s$	$\mathcal{Y}_u$	$\mathcal{X}$	$\mathcal{Y}_s$	$\mathcal{Y}_u$	$\mathcal{X}$
MIT-States	115	245	1262	30338	300	300	10420	400	400	12995
UT-Zappos	16	12	83	22998	15	15	3214	18	18	2914
C-GQA	413	674	5592	27000	1252	1040	7000	888	923	5000
VAW-CZSL	440	541	11175	72203	2121	2322	9524	2449	2470	10856

Datasets. We evaluated the model's performance on four datasets: UT-Zappos [25], MIT-States [6], C-GQA [15], and VAW-CZSL[22]. UT-Zappos is a large shoe dataset consisting of 16 attributes and 12 objects. MIT-States is a diverse collection of everyday objects, featuring 115 attributes and 245 objects. C-GQA is the largest dataset for the CZSL task, derived from the GQA dataset [5], containing 453 attributes and 870 objects. VAW-CZSL is a new, large-scale real-world attribute dataset specifically designed for Compositional Zero-Shot Learning (CZSL). It contains a diverse collection of complex attribute-object compositions, with 440 attributes and 541 objects, reflecting realistic visual scenarios. We followed the dataset split standards from previous studies[16, 17] and the statistics are provided in Tab. S1.

Implementation Details. To ensure fairness, we adopt the parameter settings established by previous research, utilizing the pre-trained CLIP ViT-L/14 model [20] as our image/text encoder. For the **AD-CA** and **OD-CA** modules, a single layer of Cross-Attention is used. During training, we use the Adam optimizer in conjunction with a StepLR learning rate scheduler, where the learning rate decays by a factor of 0.5 every 3 epochs. For the UT-Zappos and MIT-States datasets, the learning rate is 5×10^{-4} and weight decay is 1×10^{-5}; for the C-GQA and VAW-CZSL dataset, the learning rate is 5×10^{-5} with the same weight decay. Training is conducted for 20 epochs in total. All training and testing are conducted on NVIDIA A800 GPUs.

2. More Quantitative Results

We further report more comprehensive comparison results in Tab. S2 and Tab. S3, covering both ResNet18-based methods, including LE+ [14], TMN [18], SymNet [12], CompCos [8], Co-CGE [21], SCEN [10], CANet [23], and CoT [9], as well as CLIP-based approaches such as CLIP [20], CoOp [26], CSP [17], PCVL [24], DFSP [13], DLM [3], ProLT [7], PLID [1], CDS-CZSL [11], and Trokia [4].

3. Parameter Sensitivity Analysis

We conducted hyperparameter analysis of the loss function on the UT-Zappos dataset. To ensure balanced learning between textual prototypes and visual proxies, we set the same hyperparameter α for both L_t and L_v, while the hyperparameter for L_{kl} was set to β. For the three-branch prediction, the attribute and object branches share the same hyperparameter γ_{ao}, while the composition branch uses the hyperparameter γ_c:

$$\mathcal{L} = \alpha(\mathcal{L}_t + \mathcal{L}_v) + \beta\mathcal{L}_{kl} \tag{1}$$

$$\mathcal{L}_t + \mathcal{L}_v = \gamma_{ao}(\mathcal{L}_t^a + \mathcal{L}_v^a + \mathcal{L}_t^o + \mathcal{L}_v^o) + \gamma_c(\mathcal{L}_t^c + \mathcal{L}_v^c) \tag{2}$$

To evaluate the robustness of our model, we varied the parameters α, β, γ_{ao}, and γ_c within the range $\{0.1, 0.5, 1, 5, 10\}$. The results show that all hyperparameters achieve optimal performance when set to 1, suggesting that textual prototypes and visual proxies complement each other and reach optimal performance when balanced. Despite fluctuations in the hyperparameters within a certain range, our model remains stable and effective, with accuracy ranging from 45.8% to 47.9%, significantly surpassing the state-of-the-art method Troika [4], as shown in Fig. S1. This minimal variation, despite substantial parameter changes, demonstrates the stability and robustness of our model.

4. Additional Ablation Study

We evaluate our model's performance on the UT-Zappos[25] and MIT-States[6] datasets by initializing the visual proxies with text feature derived from various pre-trained language models, as summarized in Tab. S4. Specifically, we examine the effects of initializing with text features from CLIP[20], BERT [2], GPT[19], and random initialization. The experimental results indicate that using CLIP text features to initialize the visual proxies achieves the best performance.

This outcome can be attributed to the unique properties of CLIP. As a vision-language model trained on large-scale paired image-text datasets, CLIP has learned a strong correspondence between visual and textual representations. This cross-modal alignment allows its text features to serve as an effective starting point for visual prototypes, facilitating seamless integration with visual features.

In contrast, models like BERT and GPT are pre-trained exclusively on natural language tasks. While they provide semantically rich text features, their lack of alignment with the visual modality limits their effectiveness for initializing visual prototypes. Additionally, random initialization

Table S2. The experimental results for Without-CLIP and With-CLIP methods under the closed-world setting. The best performances are highlighted in bold.

Method	**Venue**	C-GQA				UT-Zappos				MIT-States			
		S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
Without CLIP													
LE+[14]	CVPR'17	18.1	5.6	6.1	0.8	53.0	61.9	41.0	25.7	15.0	20.1	10.7	2.0
TMN[18]	ICCV'19	23.1	6.5	7.5	1.1	58.7	60.0	45.0	29.3	20.2	20.1	13.0	2.9
SymNet[12]	CVPR'20	26.8	10.3	11.0	2.1	49.8	57.4	40.4	23.4	24.2	25.2	16.1	3.0
CompCos[8]	CVPR'21	28.1	11.2	12.4	2.6	59.8	62.5	43.1	28.1	25.3	24.6	16.4	4.5
Co-CGE[21]	TPAMI'22	28.1	11.9	12.7	2.8	58.2	63.3	44.1	26.1	27.8	25.2	17.5	5.1
SCEN[10]	CVPR'22	29.3	11.9	12.7	2.8	63.5	63.1	47.8	29.1	29.9	25.2	18.4	5.3
CANet[23]	CVPR'23	30.0	13.2	14.5	3.3	61.0	66.3	47.3	33.1	29.0	26.2	17.9	5.4
CoT[9]	ICCV'23	33.1	16.6	16.6	4.5	-	-	-	-	30.8	26.8	19.6	6.2
With CLIP													
CLIP[20]	ICML'21	7.5	25.0	8.6	1.4	15.8	49.1	15.6	5.0	30.2	46.0	26.1	11.0
CoOp[26]	IJCV'22	20.5	26.8	17.1	4.4	52.1	49.3	34.6	18.8	34.4	47.6	29.8	13.5
CSP[17]	ICLR'23	28.8	26.8	20.5	6.2	64.2	66.2	46.6	33.0	46.6	49.9	36.3	19.4
PCVL[24]	arXiv'22	-	-	-	-	64.4	64.0	46.1	32.2	48.5	47.2	35.3	18.3
DFSP(i2t)[13]	CVPR'23	35.6	29.3	24.3	8.7	64.2	66.4	45.1	32.1	47.4	52.4	37.2	20.7
DFSP(BiF)[13]	CVPR'23	36.5	32.0	26.2	9.9	63.3	69.2	47.1	33.5	47.1	52.8	37.7	20.8
DFSP(t2i)[13]	CVPR'23	38.2	32.0	27.1	10.5	66.7	71.7	47.2	36.0	46.9	52.0	37.3	20.6
DLM[3]	AAAI'24	32.4	28.5	21.9	7.3	67.1	72.5	52.0	39.6	46.3	49.8	37.4	20.0
ProLT[7]	AAAI'24	39.5	32.9	27.7	11.0	66.0	70.1	49.4	36.1	49.1	51.0	38.2	21.1
PLID[1]	ECCV'24	38.8	33.0	27.9	11.0	67.3	68.8	52.4	38.7	49.7	52.4	39.0	22.1
CDS-CZSL[11]	CVPR'24	38.3	34.2	28.1	11.1	63.9	74.8	52.7	39.5	50.3	52.9	39.2	22.4
Troika[4]	CVPR'24	41.0	35.7	29.4	12.4	66.8	73.8	54.6	41.7	49.0	**53.0**	39.3	22.1
VP-CMJL(Ours)		**46.0**	**40.2**	**34.9**	**16.3**	**71.9**	**76.3**	**58.5**	**47.9**	**51.8**	52.6	**40.4**	**23.3**

Table S3. The experimental results for Without-CLIP and With-CLIP methods under the open-world setting. The best performances are highlighted in bold.

Method	**Venue**	C-GQA				UT-Zappos				MIT-States			
		S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
Without CLIP													
LE+[14]	CVPR'17	19.2	0.7	1.0	0.1	60.4	36.5	30.5	16.3	14.2	2.5	2.7	0.3
TMN[18]	ICCV'19	-	-	-	-	55.9	18.1	21.7	8.4	12.6	0.9	1.2	0.1
SymNet[12]	CVPR'20	26.7	2.2	3.3	0.4	53.3	44.6	34.5	18.5	21.4	7.0	5.8	0.8
CompCos[8]	CVPR'21	28.4	1.8	2.8	0.4	59.3	46.8	36.9	21.3	25.4	10.0	8.9	1.6
Co-CGE[21]	TPAMI'22	28.7	1.6	2.6	0.4	60.1	44.3	38.1	21.3	26.4	10.4	10.1	2.0
With CLIP													
CLIP[20]	ICML'21	7.5	4.6	4.0	0.3	15.7	20.6	11.2	2.2	30.1	14.3	12.8	3.0
CoOp[26]	IJCV'22	21.0	4.6	5.5	0.7	52.1	31.5	28.9	13.2	34.6	9.3	12.3	2.8
CSP[17]	ICLR'23	28.7	5.2	6.9	1.2	64.1	44.1	38.9	22.7	46.3	15.7	17.4	5.7
PCVL[24]	arXiv'22	-	-	-	-	64.6	44.0	37.1	21.6	48.5	16.0	17.7	6.1
DFSP(i2t)[13]	CVPR'23	35.6	5.6	9.0	1.9	64.3	53.8	41.2	26.4	47.2	18.2	19.1	6.7
DFSP(BiF)[13]	CVPR'23	36.5	7.6	10.6	2.4	63.5	57.2	42.7	27.6	47.1	18.1	19.2	6.7
DFSP(t2i)[13]	CVPR'23	38.2	7.2	10.4	2.4	66.8	60.0	44.0	30.3	47.5	18.5	19.3	6.8
PLID[1]	ECCV'24	39.1	7.5	10.6	2.5	67.6	55.5	46.6	30.8	49.1	18.7	20.4	7.3
CDS-CZSL[11]	CVPR'24	37.6	8.2	11.6	2.7	64.7	61.3	48.2	32.3	49.4	**21.8**	**22.1**	**8.5**
Troika[4]	CVPR'24	40.8	7.9	10.9	2.7	66.4	61.2	47.8	33.0	48.8	18.4	20.1	7.2
VP-CMJL(Ours)		**46.0**	**11.5**	**15.5**	**4.6**	**71.9**	**66.6**	**54.5**	**41.4**	**51.8**	19.9	22.0	8.3

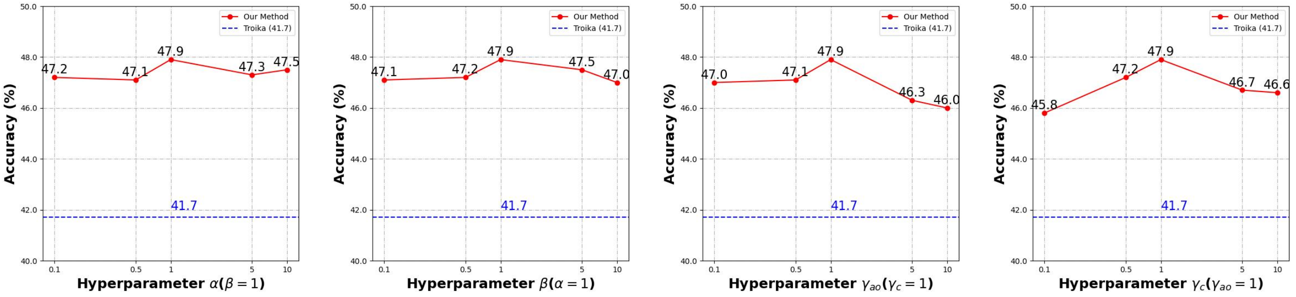
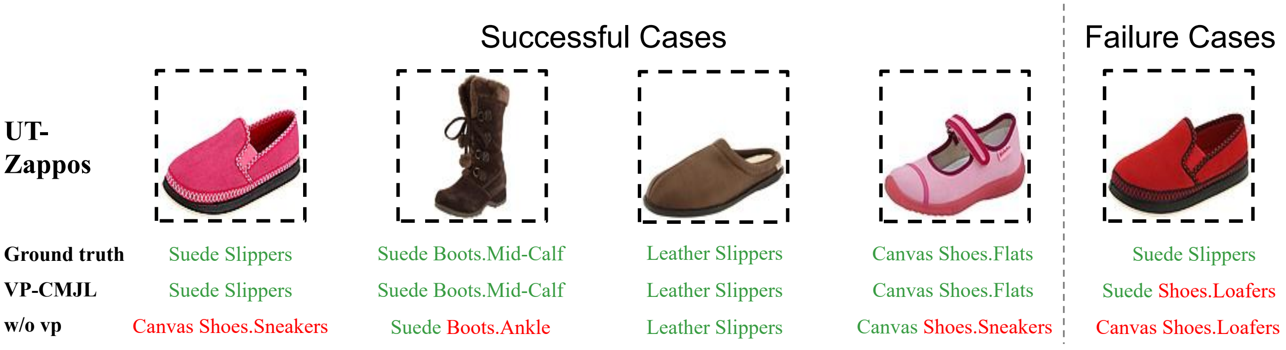


Figure S1. Sensitivity analysis on loss weighting coefficients α, β γ_{ao} and γ_c on the UT-Zappos.



Ground truth	Suede Slippers	Suede Boots.Mid-Calf	Leather Slippers	Canvas Shoes.Flats	Suede Slippers
VP-CMJL	Suede Slippers	Suede Boots.Mid-Calf	Leather Slippers	Canvas Shoes.Flats	Suede Shoes.Loafers
w/o vp	Canvas Shoes.Sneakers	Suede Boots.Ankle	Leather Slippers	Canvas Shoes.Sneakers	Canvas Shoes.Loafers

Figure S2. Qualitative results on UT-Zappos Dataset. The term 'w/o vp' refers to the text-prototype-based method, while the green font indicates correct labels and the red font indicates incorrect labels.

Table S4. Results on UT-Zappos and MIT-States datasets visual proxies with different initializations.

Model	UT-Zappos				MIT-States			
	S	U	HM	AUC	S	U	HM	AUC
CLIP[20]	**71.9**	**76.3**	**58.5**	**47.9**	**51.8**	**52.6**	**40.4**	**23.3**
BERT[2]	66.2	74.5	56.9	42.6	51.8	51.1	39.7	22.6
GPT[19]	60.31	72.24	52.0	37.0	50.8	51.3	38.4	21.8
Random	61.58	71.76	52.13	37.52	49.5	51.55	37.88	21.3

introduces significant uncertainty during training, making it more difficult for the model to converge quickly and learn optimal visual proxies.

We also observe that datasets with more similar compositions, such as UT-Zappos, are more sensitive to the initialization of visual proxies. These datasets require fine-grained visual proxies to accurately capture distinguishing features, amplifying the importance of effective proxies initialization.

Therefore, we adopt CLIP text features for initializing visual proxies to accelerate convergence and facilitate the learning of more effective and precise visual representations.

5. Additional Qualitative Visualization

To provide a more intuitive demonstration of the effectiveness of dual-modal prototypes, we visualize the image feature clustering performance and selected cases on the UT-Zappos dataset.

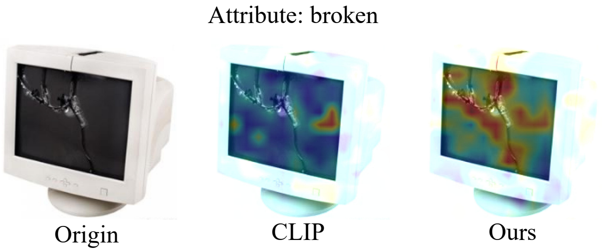


Figure S3. Comparison of Image Feature Clustering Performance between Baseline and Our Model on the UT-Zappos Dataset.

5.1. Case Study Analysis

We further visualize the qualitative results of the model on the UT-Zappos dataset in Fig. S2. Specifically, we present both successful and failure cases of the proposed **VP-CMJL** model, along with those from the text-prototype-based method, denoted as 'w/o vp'. The results clearly show that **VP-CMJL** can accurately distinguish between visually similar compositions, such as 'Suede Boots.Mid-Calf' and 'Suede Boots.Ankle', whereas the text-prototype-based method struggles to differentiate compositions with similar visual appearances. This demonstrates that **VP-CMJL** effectively learns fine-grained compositional features. In failure cases, although the model does not always correctly identify the complete composition, it often successfully classifies at least one of the primitives. Furthermore, we employ Grad-CAM to visualize the model's ability to capture fine-grained classification cues. As shown in Fig. S3, for fine-grained attributes such as 'broken', our model is able to localize more precise visual cues compared to CLIP, thereby enhancing category discriminability.

5.2. Visualization of Image Feature Clustering

We first generate a t-SNE visualization of image features for six categories from the UT-Zappos dataset, as depicted in Fig. S4. Compared to the baseline, our model significantly reduces intra-class distances and increases inter-class distances. This demonstrates that the proposed visual prototypes effectively enhance feature learning within the visual modality.

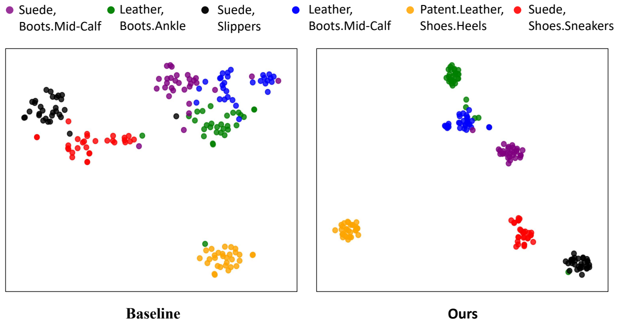


Figure S4. Comparison of Image Feature Clustering Performance between Baseline and Our Model on the UT-Zappos Dataset.

References

[1] Wentao Bao, Lichang Chen, Heng Huang, and Yu Kong. Prompting language-informed distribution for compositional zero-shot learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1, 2

[2] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 3

[3] Xiaoming Hu and Zilei Wang. A dynamic learning method towards realistic compositional zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2265–2273, 2024. 1, 2

[4] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24005–24014, 2024. 1, 2

[5] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1

[6] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391, 2015. 1

[7] Chenyi Jiang and Haofeng Zhang. Revealing the proximate long-tail distribution in compositional zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2498–2506, 2024. 1, 2

[8] Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9336–9345, 2022. 1, 2

[9] Hanjae Kim, Jiyoung Lee, Seongheon Park, and Kwanghoon Sohn. Hierarchical visual primitive experts for compositional zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5675–5685, 2023. 1, 2

[10] Xiangyu Li, Xu Yang, Kun Wei, Cheng Deng, and Muli Yang. Siamese contrastive embedding network for compositional zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9326–9335, 2022. 1, 2

[11] Yun Li, Zhe Liu, Hang Chen, and Lina Yao. Context-based and diversity-driven specificity in compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17037–17046, 2024. 1, 2

[12] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11316–11325, 2020. 1, 2

[13] Xiaocheng Lu, Song Guo, Ziming Liu, and Jingcai Guo. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23560–23569, 2023. 1, 2

[14] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1792–1801, 2017. 1, 2

[15] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 953–962, 2021. 1

[16] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 953–962, 2021. 1

[17] Peilin Yu Nihal V. Nayak and Stephen H. Bach. Learning to compose soft prompts for compositional zero-shot learning. In *In Proceedings of the International Conference on Learning Representations*, 2023. 1, 2

[18] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3593–3602, 2019. 1, 2

[19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1, 3

[20] 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, 3

[21] Frank Ruis, Gertjan Burghouts, and Doina Bucur. Independent prototype propagation for zero-shot compositionality. *Advances in Neural Information Processing Systems*, 34: 10641–10653, 2021. 1, 2

[22] Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Disentangling visual embeddings for attributes and objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13658–13667, 2022. 1

[23] Qingsheng Wang, Lingqiao Liu, Chenchen Jing, Hao Chen, Guoqiang Liang, Peng Wang, and Chunhua Shen. Learning conditional attributes for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11197–11206, 2023. 1, 2

[24] Guangyue Xu, Parisa Kordjamshidi, and Joyce Chai. Prompting large pre-trained vision-language models for compositional concept learning. *arXiv preprint arXiv:2211.05077*, 2022. 1, 2

[25] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 192–199, 2014. 1

[26] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, pages 2337–2348, 2022. 1, 2