Appendix of "Tuning Multi-mode Token-level Prompt Alignment across Modalities"

Anonymous Author(s)
Affiliation
Address
email

A Discussion

We in this paper propose ALIGN, a unified framework for multi-modal prompt tuning, where multi-mode modality-specific prompts are learned via the token-level alignment strategy. Moving beyond the single-model methods, which focus on textual prompt tuning or visual prompt tuning, ALIGN allows one to learn textual and visual prompts simultaneously, resulting in better representations in the shared vision-text embedding space. Compared to recent multi-modal methods, such as UPT [1] and MaPLe [2], ALIGN prefers to learn multi-mode prompts to capture diverse class attributes and develop the token-level alignment for fine-grained comparisons. This provides ALIGN with an efficient tool to calculate the similarity between prompts. We find that many previous works can be merged into our ALIGN framework with special hypermeter settings. We summarize this relationship at Table. 1. The N/A in Table. 1 means that PLOT calculates the similarity between the prompt-level label embeddings and the visual patch embeddings, which is not the case in ALIGN, where we calculate the similarity of prompt-level OT between textual label embeddings and visual image embeddings. and calculate the similarity of token-level OT between token embeddings and patch embeddings.

Table 1: Most previous works can be merged into our ALIGN framework. M: Number of visual prompts. N: Number of textual prompts. β: Weight of token-level OT in Eq.6 in the manuscript.

Methods	Type	M	N	β
CoOp [3]	Textal Prompt Tuning	0	1	0
VPT [4]	Visual Prompt Tuning	1	0	0
PLOT [5]	Textual Prompt Tuning	0	≥ 1	N/A
UPT [1]	Multi-modal Prompt Tuning	1	1	0
MaPLe [2]	Multi-modal Prompt Tuning	1	1	0
ALIGN(Ours)	Multi-modal Prompt Tuning	≥ 0	≥ 0	≥ 0

B Data statistics and Hyperparameter setting

We thoroughly evaluate our proposed ALIGN framework across four distinct tasks: few-shot recognition, base-to-new generalization, cross-dataset transfer, and cross-domain generalization. These extensive experiments are conducted on a diverse set of fifty commonly used vision datasets, covering various contexts. These datasets include ImageNet [6] and Caltech101 [7] for generic image classification, OxfordPets [8], StanfordCars [9], Flowers102 [10], Food101 [11], and FGVCAircraft [12] for fine-grained image recognition, SUN397 [13] for scene recognition, UCF101 [14] for action recognition, DTD [15] for texture classification, and EuroSAT [16] for satellite imagery recognition. In the case of the cross-domain generalization task, our model is trained on ImageNet and subsequently tested on ImageNetV2 [17], ImageNet-Sketch [18], ImageNet-A [19], and ImageNet-R [20]. We summarize data statistics at Table. B. 1

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

The evaluation pipeline for each task follows the approach employed by previous works [3, 2]. The specific details of this pipeline are summarized below:

Few-shot Recognition. To evaluate the efficiency of the proposed ALIGN on the few-shot case, we follow CoOp [3], and first partition the dataset into base and novel sets. Those two sets share the same categories. Models are trained on the base set using a variety of shot settings, including 1, 2, 4, 8, and 16 shots per class, and then tested on the full novel set. The accuracy scores are reported to compare the performance. The training epoch is set as 10 for 1, 2, and 4 shots and 40 for 8 and 16 shots.

Base-to-New Generalization. To show the Generalizability of unseen categories, we first divide the dataset into two separate subsets: the base subset and the new subset. Importantly, these subsets do not share the same categories. The base subset contains a specific set of categories used for model training, while the new subset consists of previously unseen categories that the model has not been exposed to during training. Besides reporting the accuracy score on base and novel sets, we also calculate the harmonic mean $\mathrm{H} = (2 \times \mathrm{Base} \times \mathrm{New})/(\mathrm{Base} + \mathrm{New})$, which acts as a trade-off between Base and New, providing a comprehensive measure of overall model performance. The training epoch is set as 8.

Cross-Dataset Transfer. To determine the transferability of our model across different datasets, we first train our model on the source dataset (ImageNet) and then evaluate it on 10 different target datasets. The training epoch is set as 2 and the learning rate is set as 0.0026.

Cross-Domain Generalization. To determine the robustness of our model on the distribution-shift setting, we trained our model on the source dataset (ImageNet) and then assess it on 4 domain-shifted datasets, including ImageNetV2, ImageNet-Sketch, ImageNet-A, and ImageNet-R. The training epoch is set as 2 and the learning rate is set as 0.0026.

The other training hyperparameters in the previous experiments are set according to MaPLe [2], which are detailed listed at Table B. 2.

Table B. 1: Statistics of the used 15 datasets. N/A denotes that we do not use the corresponding training or validation sets.

Dataset	Domains	**#Classes**	**#Train**	**#Val**	**#Test**
ImageNet	generic object	1000	1.28M	N/A	50,000
Caltech101	generic object	100	4,128	1,649	2,465
OxfordPets	fine-grained object	37	2,944	736	3,669
StanfordCars	fine-grained object	196	6,509	1,635	8,041
Flowers102	fine-grained object	102	4,093	1,633	2,463
Food101	fine-grained object	101	50,500	20,200	30,300
FDVCAircraft	fine-grained object	100	3,334	3,333	3,333
SUN397	scene recognition	397	15,880	3,970	19,850
UCF101	action recognition	101	7,639	1,808	3,783
DTD	texture recognition	47	2,820	1,128	1,692
EuroSAT	satellite object	10	13,500	5,400	8,100
ImageNetV2	generic object	1000	N/A	N/A	10,000
ImageNet-Sketch	sketch object	1000	N/A	N/A	50,889
ImageNet-A	generic object	200	N/A	N/A	7,500
ImageNet-R	generic object	200	N/A	N/A	30,000

C Training Algorithm

Given the training datasets $\mathcal{D} = \{\mathcal{X}_i, y_{\mathcal{X}_i}\}_{i=1}^{N_\mathcal{D}}$, our method aims to learn M visual and N textual prompts simultaneously. All parameters in ALIGN are optimized by minimizing the cross-entropy loss end-to-end. We summarize the training algorithm at Algorithm. 1.

Table B. 2: Hyperparameter setting used in the previous experiments.

Hyperparameters	**Values**
Batch Size	4
Input Size	224×224
Input Interpolation	"Bicubic"
Input Pixel Mean	$[0.48145466, 0.4578275, 0.40821073]$
Input Pixel STD	$[0.26862954, 0.26130258, 0.27577711]$
Transforms	["random resized crop", "random filp", "normalize"]
Optimizer	SGD
Learning Rate	0.0035
LR Scheduler	"cosine"
Warmup Epoch	1
Warmup Type	"constant"
Warmup LR	$1e$-5
Backbone	ViT-B/16
Number of Textual Prompts	4
Number of Visual Prompts	4
Learnable Prompt Length	2
Fixed Prompt Length	2
weight of token-level cost	1
weight of regularization in OT	0.1
Prompt Initialization	"a photo of a"
Precision	"fp16"

Algorithm 1 Training algorithm of ALIGN.

Input: Training dataset $\mathcal{D}$, a pre-trained vision-language model, class name set, number of visual prompts M, number of textual prompts N, and the training epoch.
Output: The learned ALIGN, which discovers multi-modal multi-mode prompts for downstream tasks.
Initialize: The M and N multi-modal prompt embeddings.
Preprocess: Built $N \times K$ textual token inputs according to Sec 2.1 in the manuscript.
for iter = 1,2,3,... **do**
1. Feed the textual input into the text encoder g and collect the outputs with the corresponding prompt-level representation $\{\boldsymbol{h}_k^n\}_{k=1,n=1}^{K,N}$ and token embeddings $\{\boldsymbol{s}_k^n\}_{k=1,n=1}^{K,N}$, where each $\boldsymbol{s}_k^n$ is the output token embeddings of n-th prompt of k-th label with length $b + k_l$.
2. Sample a batch of J images. Built $N \times B$ visual patch inputs according to Sec2.1 in the manuscript. Feed the visual input into the visual encoder f and collect the outputs with the corresponding prompt-level representation $\{\boldsymbol{z}_j^m\}_{j=1,m=1}^{J,M}$ and patch embeddings $\{\boldsymbol{r}_j^m\}_{j=1,m=1}^{J,M}$, where each $\boldsymbol{r}_j^m$ denotes the output patch embeddings of m-th prompt of j-th image with length $b + O$.
Two-level OT
3. Calculate the token-level OT distance between each image and each label in Eq.5 with the collected patch set and token set.
4. Calculate the cost matrix in prompt-level OT according to Eq.6, and then get the prompt-level OT distance in Eq.4.
Compute the cross-entropy loss L with the obtained prompt-level OT distance according to Eq.8 and update all learnable parameters by minimizing L with the stochastic gradient descent algorithm.
end for

D Additional Results

We in this section report additional results of other datasets on the few-shot task and conduct the ablation studies on the prompt and token-level OT.

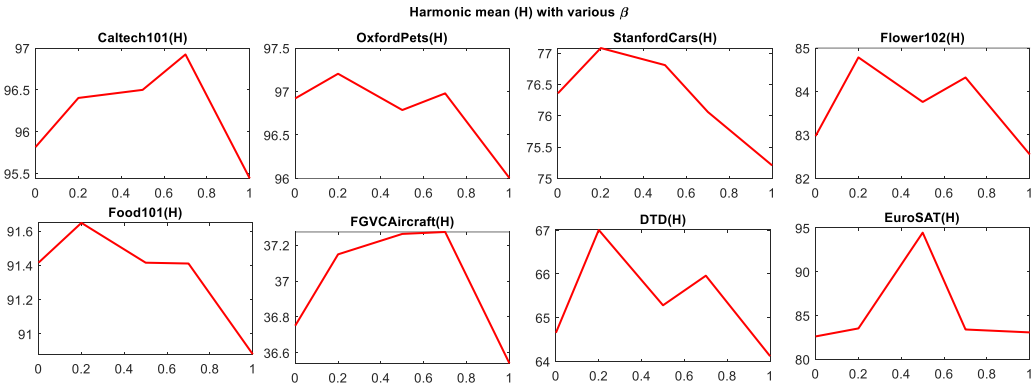

Figure D1: Harmonic mean (H) results of ALIGN on Base-to-New task under different β.

D.1 Few-shot Results

We report the numerical results of various methods on 11 datasets at Table. D. 1.. From the results, we find that our method ALIGN outperforms baselines in most cases, which demonstrates the efficiency of the token-level prompt alignment module.

D.2 Ablation studies

Recall that the proposed model consists of the prompt-level and token-level OT, which align the textual and visual modalities from hierarchical semantics. In the previous experiment, we view the prompt-level and token-level OT equally and set the hyperparameter weight $\beta = 1$ in Eq.6 in the manuscript. Here want to analyze how those two OTs affect the model performance. To this end, we rewrite the cost matrix in Eq.6 in the manuscript as:

$$C_{mn} = (1-\beta)(1-\text{sim}(\boldsymbol{z}^m, \boldsymbol{h}^n)) + \beta d_{\text{OT}}^{\lambda}(\boldsymbol{x}_m, \boldsymbol{y}_n; \hat{\mathbf{C}}^{mn}). \tag{1}$$

Note that $\beta = 0$ and $\beta = 1$ denote two of our variants, where the former denotes only prompt-OT works and the latter means we only focus on token-level similarity. We report the ablation results of ALIGN on Base-to-New tasks under various settings, *e.g.*, $\beta = [0, 0.2, 0.5, 0.7, 1.0]$ at Fig. D1. We have the following interesting findings: 1) The combined ALIGN works better than each of them; 2) After finetuning β for each dataset, one can obtain better results than the reported values in our paper.

Table D. 1.: The few-shot results of various methods on 11 datasets. We report mean value over 3 different seeds. The best results are **highlighted**.

Dataset	Methods	1 shot	2 shots	4 shots	8 shots	16 shots
Caltech101	CoOp	92.4	93.2	93.5	94.0	94.8
	PLOT	88.40	89.95	91.50	93.00	93.24
	UPT	93.66	94.17	94.09	95.04	95.95
	MaPLe	91.73	93.33	94.23	94.43	95.26
	ALIGN	**93.97**	**94.13**	**95.00**	**95.43**	**96.00**
DTD	CoOp	48.4	51.5	59.2	64.4	69.5
	PLOT	51.90	55.95	58.24	65.50	70.52
	UPT	45.01	52.97	60.74	65.44	70.62
	MaPLe	51.16	54.70	61.63	65.63	70.60
	ALIGN	**54.07**	**56.53**	**63.3**	**67.6**	**71.40**
EuroSAT	CoOp	51.8	60.9	69.0	76.0	84.1
	PLOT	60.10	68.45	72.97	79.84	83.12
	UPT	66.46	69.07	75.36	85.62	90.77
	MaPLe	**66.67**	**79.26**	**84.25**	**89.96**	**92.14**
	ALIGN	53.23	71.43	80.93	85.97	90.77
FGVCAircraft	CoOp	24.2	25.8	27.9	32.7	34.2
	PLOT	21.50	21.71	23.96	27.02	30.24
	UPT	28.43	29.91	33.34	39.50	46.61
	MaPLe	26.64	27.86	33.56	40.66	49.93
	ALIGN	**29.57**	**31.63**	**34.03**	**40.95**	**49.99**
Flowers102	CoOp	72.9	80.4	85.7	92.3	96.2
	PLOT	70.00	81.34	88.29	92.84	95.10
	UPT	74.97	81.81	91.90	95.17	**97.41**
	MaPLe	80.24	88.14	90.07	95.10	96.34
	ALIGN	**81.33**	**88.77**	**92.53**	**95.43**	96.57
FOOD101	CoOp	81.6	80.9	81.5	82.4	84.9
	PLOT	69.10	72.89	74.89	76.70	77.87
	UPT	84.21	85.01	85.34	86.16	86.83
	MaPLe	78.73	77.30	79.03	80.10	82.43
	ALIGN	**85.29**	**86.05**	**86.66**	**86.74**	**86.90**
ImageNet	CoOp	68.07	69.26	69.60	70.35	71.53
	PLOT	67.51	68.80	70.00	70.21	71.40
	UPT	69.55	69.88	70.28	71.58	72.64
	MaPLe	69.56	69.94	70.65	**71.80**	**72.74**
	ALIGN	**69.80**	**70.02**	**70.84**	71.77	72.45
OxfordPets	CoOp	90.0	89.8	92.3	92.0	92.1
	PLOT	83.21	85.77	86.02	89.13	89.95
	UPT	82.93	85.40	85.97	87.40	88.10
	MaPLe	89.80	86.76	90.76	90.23	91.30
	ALIGN	**91.36**	**91.93**	**93.4**	**93.67**	**94.17**
StanfordCars	CoOp	66.4	69.2	70.1	72.8	75.2
	PLOT	46.20	51.67	54.35	60.52	65.32
	UPT	67.60	69.57	75.88	80.19	84.17
	MaPLe	65.96	69.10	75.73	79.76	85.36
	ALIGN	**68.27**	**72.84**	**76.58**	**81.65**	**86.75**
SUN397	CoOp	65.2	66.6	68.1	70.5	73.2
	PLOT	55.33	60.02	63.21	66.02	67.98
	UPT	68.84	69.76	**72.12**	74.00	75.90
	MaPLe	61.73	63.23	67.60	69.13	73.00
	ALIGN	**69.14**	**69.98**	71.88	**74.15**	**76.57**
UCF101	CoOp	70.7	73.8	76.6	79.6	80.4
	PLOT	51.42	54.89	61.23	67.45	70.85
	UPT	71.98	74.93	77.49	80.91	83.86
	MaPLe	73.23	73.00	77.45	81.2	84.67
	ALIGN	**74.42**	**75.87**	**80.18**	**81.99**	**95.69**

References

[1] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *CoRR*, abs/2210.07225, 2022.

[2] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. *arXiv preprint arXiv:2210.03117*, 2022.

[3] 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.

[4] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIII*. Springer, 2022.

[5] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. PLOT: Prompt learning with optimal transport for vision-language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=zqwryBoXYnh`.

[6] 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. Ieee, 2009.

[7] 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.

[8] 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.

[9] 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.

[10] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

[11] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014.

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

[13] 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. IEEE, 2010.

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

[15] 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.

[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.

[17] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

[18] 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.

[19] 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.

[20] 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.