Supplementary Material of "FedMPT: Federated Multi-label Prompt Tuning of Vision-Language Models"

Anonymous CVPR submission

Paper ID *****

A. More Experiments

A.1. More Ablation Studies on Participation Rate

Table 1 presents extended ablation studies on all baselines. FedMPT consistently outperforms all state-of-the-art methods by substantial margins, achieving gains of 2.22% mAP, 2.88% CF1, and 3.26% OF1. Notably, methods relying more heavily on visual adaptation (e.g., FedMVP and FedMaPLe) exhibit significantly higher performance variance as the client participation rate decreases. This can be attributed to their local models' heightened susceptibility to overfitting client-specific data; when aggregated under low participation rates, the global model is disproportionately influenced by these overfitted local models, making it more vulnerable to heterogeneity and distribution shifts.

A.2. Experiments on ZSL and GZSL Benchmarks

Following RAM [15], we conduct more experiments on two other benchmarks [5, 14]: the Federated Zero-shot Generalization Benchmark (FZSL) and the Federated Generalized Zero-shot Generalization Benchmark (FGZSL). These two benchmarks evaluate the model's robustness to unseen classes, which is a comparably harsher setting.

Benchmarks Overview. The Federated Zero-shot Generalization Benchmark (ZSL) first splits all classes into seen and unseen classes, then performs clustering on training samples and sends each cluster to a client as its private data. Local models are trained on their private data, with only the seen classes annotated. The global model is evaluated on the test data, with only the unseen classes considered in all of the metrics. The Federated Generalized Zero-shot Generalization Benchmark (FGZSL) is similar, but the global model is evaluated on the test data, with both seen and unseen classes considered in all of the metrics. COCO2014 [8] and NUS-Wide [1] are employed for the above two benchmarks. For COCO2014, the dataset is split into 48 seen classes and 17 unseen classes. NUS-WIDE is split into 925 seen classes and 81 unseen classes.

Results on FZSL Benchmark. As shown in Table 2, all methods exhibit significant performance degradation on the challenging COCO2014 benchmark. For instance, FedMaPLe achieves only 2.63 mAP, while FedMVP reaches 6.53 mAP. These results underscore the particular difficulty of achieving robust class-level generalization in federated learning environments. In contrast, FedMPT substantially outperforms all SOTA methods across both datasets, demonstrating superior generalization and robustness.

Results on FGZSL Benchmark. As shown in Table 5, the performance gain of FedMPT is consistent, as it outperforms existing SOTAs by 3.79 mAP and 3.91 mAP on COCO2014 and NUS-Wide, respectively. This result demonstrates FedMPT's substantial generalization capabilities across both seen and unseen categories.

A.3. Experiments of convergence speed and statistical significance

The experiment results of convergence speed and statistical significance are shown in Table 3 and Table 4.

B. Ablation Studies on Hyper-parameters

$\gamma+$ and $\gamma-$. We report the experiment results in Figure 1 (left). We find that both excessively small and large values lead to performance degradation, potentially due to under-constrained/over-constrained optimization for easy-negative samples. Based on experimental results, we selected the values (1,2) for ($\gamma+$, $\gamma-$).

The threshold c. This coefficient controls the clip threshold, where logits below it are clamped to 0. We alter c from 0.01 to 0.2 and report the results in 1 (right). We observed that both excessively small and large values lead to performance degradation, likely due to excessive influence from low-confidence tail classes and over-clipped negative logits, respectively. Based on experimental results, we set this parameter to 0.05.

Table 1. **More ablation studies of FedMPT and other methods on the participation rate** . We report the mAP, CF1 and OF1 with the part-annotation setting `Mask` varying from 10% to 90%. The best results are marked with **bold**.

Method	Venues	Pat. rat.=0.1			Pat. rat.=0.3			Pat. rat.=0.5			Pat. rat.=0.7			Pat. rat.=0.9			Avg		
		mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1
Fed-DualCoOp	NeurIPS'22	85.92	79.88	77.74	83.59	77.57	76.36	84.70	77.85	75.73	84.27	76.95	78.13	85.44	78.27	76.98	84.78	78.10	76.99
Fed-SCPNet	CVPR'23	79.67	68.96	72.86	76.38	69.29	69.98	77.09	71.60	70.71	78.96	70.70	72.17	80.44	77.37	74.25	78.51	71.58	71.99
Fed-MaPLe	CVPR'23	82.48	75.60	73.68	83.18	77.46	72.99	82.81	77.54	75.03	87.10	81.46	78.48	86.25	81.25	79.87	84.36	78.66	76.01
FedPGP	ICML'24	79.43	72.56	68.38	78.68	73.29	69.40	81.18	75.16	65.22	81.57	74.17	66.58	83.55	76.60	72.77	80.88	74.36	68.47
Fed-TCP	CVPR'24	79.66	73.58	74.14	81.19	75.31	72.07	81.83	76.95	70.63	82.07	76.00	69.15	83.00	77.80	79.12	81.55	75.93	73.02
FedTPG	ICLR'24	83.29	75.02	76.83	84.73	76.01	78.23	86.92	80.47	79.95	86.10	80.64	77.26	87.89	80.12	77.62	85.79	78.45	77.98
Fed-PosCoOp	WACV'25	80.33	73.63	78.62	80.39	75.49	79.69	82.65	76.04	75.27	84.02	78.97	68.99	87.31	83.29	80.57	82.94	77.48	76.63
FedAWA	CVPR'25	79.98	70.04	78.11	78.33	68.31	69.77	80.09	71.63	70.30	84.45	72.93	70.97	83.82	75.31	72.73	81.33	71.64	72.38
Fed-RAM	CVPR'25	86.91	79.09	80.05	88.53	82.13	82.56	87.20	81.45	80.40	87.64	80.70	81.84	87.29	80.85	78.29	87.51	80.84	80.63
FedMVP	ICCV'25	80.09	72.73	72.80	80.81	72.50	72.12	82.36	72.23	70.88	84.63	77.93	78.03	87.76	82.88	80.00	83.13	75.65	74.77
FedMPT	Ours	**89.25**	**82.41**	**84.78**	**89.23**	**82.97**	**84.05**	**89.96**	**83.74**	**83.25**	**90.10**	**84.51**	**83.30**	**90.15**	**84.97**	**84.05**	**89.74**	**83.72**	**83.89**
Δ Prev. Best	\	**+2.34**	**+2.53**	**+4.73**	**+0.70**	**+0.84**	**+1.49**	**+2.76**	**+2.29**	**+2.85**	**+2.46**	**+3.05**	**+1.46**	**+2.26**	**+1.68**	**+3.48**	**+2.22**	**+2.88**	**+3.26**

Table 2. **Results on the FZSL benchmark.** We report the mAP, CF1 and OF1. The best results are marked with **bold**.

Method	COCO2014			NUS-WIDE		
	mAP	CF1	OF1	mAP	CF1	OF1
Fed-DualCoOp	11.19	8.59	10.97	48.65	50.39	67.52
Fed-SCPNet	5.89	5.34	5.71	38.35	41.62	61.61
Fed-MaPLe	2.63	2.95	2.28	53.13	53.28	68.65
FedPGP	6.13	4.98	4.90	51.10	53.07	67.72
Fed-TCP	9.40	8.97	9.99	42.90	44.77	64.30
FedTPG	10.53	9.90	8.37	42.14	40.36	65.30
Fed-PosCoOp	10.68	8.22	8.17	47.44	46.03	64.47
FedAWA	14.88	10.56	8.64	44.40	47.78	64.75
Fed-RAM	13.84	8.77	9.22	50.52	42.69	66.79
FedMVP	6.53	5.73	2.09	52.73	52.61	70.67
FedMPT	**19.76**	**16.02**	**15.38**	**57.17**	**56.94**	**71.83**
Δ Prev. Best	**+4.88**	**+5.46**	**+4.41**	**+4.04**	**+3.66**	**+1.16**

Table 3. The convergence speed (x-axis: communication round).

Method | VOC	1	5	10	20	30	40	50	60	70	80	90	100
Fed-DualCoOp	59.72	76.28	80.01	81.66	82.65	83.31	83.74	84.05	84.26	84.38	84.40	84.41
FedRAM	61.64	77.50	81.11	82.29	83.65	84.52	85.04	85.31	85.55	85.67	85.68	85.68
FedMVP	53.49	76.39	78.54	80.13	81.90	82.36	83.09	83.92	84.64	85.18	85.52	85.61
FedMPT	67.39	80.11	85.76	87.32	88.63	89.20	89.71	89.98	90.00	90.03	90.04	90.04

Table 4. The effects of gating (upper) and significance test (lower).

Metric	VOC2007			COCO2014		
	mAP	CF1	OF1	mAP	CF1	OF1
no-gating's SD.	±1.12	±0.73	±1.31	±0.38	±0.40	±0.53
FedMPT's SD.	±0.27	±0.48	±0.50	±0.22	±0.29	±0.39
p-value	$1.907e^{-6}$	$1.907e^{-6}$	$1.907e^{-6}$	$1.907e^{-6}$	$1.907e^{-6}$	$1.907e^{-6}$
Cliff's Delta	1.000	1.000	1.000	1.000	0.990	1.000

Table 5. **Results on the FGZSL benchmark.** We report the mAP, CF1 and OF1. The best results are marked with **bold**.

Method	COCO2014			NUS-Wide		
	MAP	CF1	OF1	MAP	CF1	OF1
Fed-DualCoOp	52.34	44.34	64.27	50.31	52.86	66.86
Fed-SCPNet	36.07	33.31	53.00	36.50	40.76	59.23
Fed-MaPLe	52.73	42,56	64.24	52.09	53.50	69.00
FedPGP	43.76	37.54	58.88	44.46	47.59	63.76
Fed-TCP	33.94	31.33	52.42	30.05	34.36	58.21
FedTPG	41.22	38.49	53.75	32.51	36.23	56.41
Fed-PosCoOp	39.19	35.92	53.96	46.57	45.43	51.78
FedAWA	36.91	32.81	53.60	41.04	44.94	61.00
Fed-RAM	50.72	24.25	35.00	50.72	24.25	35.00
FedMVP	43.95	36.72	60.03	48.79	51.31	67.37
FedMPT	**56.52**	**47.92**	**68.09**	**56.00**	**55.83**	**72.87**
Δ Prev. Best	**+3.79**	**+3.58**	**+3.82**	**+3.91**	**+2.33**	**+3.87**

C. Introduction of datasets and baselines

Datasets: We employ VOC2007 [4], COCO2014 [8], NUS-Wide [1], Multi-Scene [6] and MLRSNet [9] in our experiments. Details are in the following:

- VOC2007 is a commonly-employed dataset in classification and object detection. It contains 9,963 real-world images annotated with 24,640 object instances across 20 different categories, including people, animals, vehicles, and household items. The dataset supports multi-label classification, detection, segmentation, and even person layout identification (predicting parts of a person). The diversity of scenes make VOC2007 a standard benchmark for evaluating object recognition algorithms.
- COCO2014 is a large-scale dataset for object detection, segmentation, and captioning. It includes over 330,000 images, more than 200,000 of which are labeled, encompassing around 1.5 million object instances. COCO2014 version covers 80 object categories (from a larger set of 91 classes) with per-instance segmentation masks, making it especially useful for precise localization. Beyond detection and segmentation, COCO2014 also supports captioning (5 captions per image) and keypoint detection (e.g., human pose), facilitating research in richer scene understanding. COCO is considered one of the most challenging and representative vision benchmarks.
- NUS-WIDE is a large-scale multi-label image dataset derived from Flickr. It comprises 269,648 images annotated with 81 ground-truth "concepts" (e.g., sky, building, person) plus up to 5,018 user-provided noisy tags. Since the original Flickr tags are noisy and incomplete, NUS-WIDE poses realistic challenges for multi-label learning, annotation, and retrieval. In addition, it provides low-level visual features for each image.

Baselines. We employed ten baselines: DualCoOp [14], SCPNet [3], PosCoOp [11], RAM [15], MaPLe [7], TCP [16], FedPGP [2], FedTPG [10], FedAWA [12] and FedMVP [13]. Details are as follows:

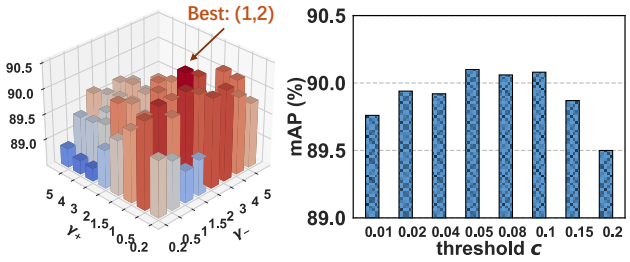

Figure 1. Ablation on **(left):** γ_- / γ_+ and **(right):** c.

- DualCoOp [14] is the first approach that leverages pre-trained VLMs (specifically, CLIP) for multi-label recognition. It introduces two prompts, named Positive Prompt and Negative Prompt, to reflect the existence and non-existence of a label.
- SCPNet [3] (Semantic Correspondence Prompt Network) proposes to extract the structured semantic prior between labels from CLIP via a structured prior prompter. It then fully explores this prior using a cross-modality prompter and a semantic association module to improve the performance of multi-label recognition with incomplete labels.
- PosCoOp [11] finds that the negative prompt in DualCoOp [14] does not necessarily need to be conditioned on class names. It leaves the negative prompts unconditionally learnable and only generates positive prompts from class names.
- RAM [15] recovers the local semantics of CLIP in a memory-efficient manner through the Ladder Local Adapter (LLA), which addresses the loss of local information caused by CLIP's global pretraining objectives. It also designs Knowledge-Constrained Optimal Transport (KCOT), formulating region-to-label matching as an optimal transport problem and integrating Label Presence Detection (LPD) and Teacher Knowledge Transfer (TKT) to suppress meaningless matching, thereby improving the performance of open-vocabulary multi-label recognition.
- MaPLe [7] introduces multi-modal prompts to both encoders. The text prompts are dynamically generated from visual prompts via a cross-modality projector.
- TCP [16] maps class-level textual knowledge into class-aware prompt tokens through the Textual Knowledge Embedding (TKE) module and then injects them into the text encoder. The algorithm optimizes the model using contrastive loss and knowledge-guided consistency loss. Notably, TKE is a plug-and-play design that can be combined with existing prompt tuning methods, achieving high performance while reducing training time.
- FedTPG [10] jointly learns a unified prompt generation network across multiple clients under the federated learning framework. This network generates context-aware prompt vectors conditioned on task-related text inputs, enabling efficient generalization to both seen and unseen classes and datasets.
- FedAWA [12] obtains client vectors by calculating the difference between client model and global model parameters. It then adaptively optimizes aggregation weights based on the alignment between these client vectors and the aggregated global vector and introduces a regularization term to ensure training stability, mitigating the problem of data heterogeneity without requiring proxy data.
- FedMVP [13] proposes to fuse the visual features of images and the textual attribute features of classes via the PromptFormer module to generate multimodal visual prompts, injects them into the vision encoder of CLIP, and trains the model by combining CLIP similarity loss and consistency loss, thereby improving the generalization ability to unseen classes and domains under the federated learning framework.

Table 6. Lists of conditions used for different datasets.

Dataset	LLM-generated conditions
VOC2007	["background", "position", "shape", "action"]
COCO2014	["background", "position", "shape", "action"]
NUS-WIDE	["color", "texture", "shape", "action"]
Multi-Scene	["color", "geometry", "shape", "contrast"]
MLRSNet	["size", "color", "shape", "texture"]

D. Condition Study

Conditions we used in the experiments. The conditions we used for 5 datasets are listed in Table 6.

Ablation study of condition number and condition combinations in LLM queries. While this factor largely depends on the LLM itself and shows considerable uncertainty, we vary it in the following manner and report the results on VOC2007. We keep the first instruction in our Chain-of-Thought mechanism unchanged, i.e., Please give a detailed description for each possible combination of the following categories in one sentence. Categories: Aeroplane, Bicycle, Bird, Boat, Bottle,..., then change the required number K (from 1 to 20) in the other instruction: Given these descriptions, Please summarize K distinct and general conditions under which true class correlations can be reliably represented.. We conduct five independent API calls and report the most frequent conditions generated by the LLM in Table 9. We observe that when instructing an LLM to generate very few conditions, the resulting conditions tend to be overly broad (e.g., "context"); conversely, requesting an excessive number of conditions yields outputs that are either overly specific or difficult to observe, such as "perspective" and "reflectance". Furthermore, we organize all obtained conditions into prompts for training and record their corresponding accuracy scores, also in Table 9. We find that using either a few broad conditions (like context,"layout") or overwhelmingly specific conditions yields

Table 7. Ablation study on different conditions.

Context	mAP	Context	mAP
null	86.48	null	86.48
+ background	88.61	+ pattern	87.23
+ position	88.98	+ anchor	87.78
+ shape	89.32	+ action	88.72
+ action	90.10	+ habits	88.59

Table 8. Ablation study on condition orders.

Condition Order	mAP
[background, position, shape, action]	90.10
[position, shape, action, background]	89.97
[shape, action, background, position]	90.02
[action, background, position, shape]	90.08
[background, shape, position, action]	90.05
[shape, position, action, background]	90.02
[position, action, background, shape]	89.89
[action, background, shape, position]	90.01

suboptimal performance. These results suggest that selecting a moderate quantity represents a reasonable trade-off. As can be observed from the table, we set the required condition number in LLM queries to 4 as a trade-off between efficiency and effectiveness.

We also conduct an experiment to discover the inherent efficiency contrasts of different conditions. Based on the conditions we used in our experiments for VOC2007, i.e., *background, position, shape, action*, we ask GPT-4o to generate some similar but comparably vague and non-representative conditions with: These are some conditions under which true class correlations in multi-label datasets can be reliably represented: *background, position, shape, action*. Now think conversely. For each given condition, please give another condition that is similar in meaning, but under which the true class correlations in multi-label datasets cannot be reliably represented. Under five independent API queries, GPT-4o mostly generates *pattern, anchor, surface, habits*, which show more incongruity and are harder to perceive. We then gradually add both kinds of conditions to prompts to discover each condition's effects. The results are in Table 7. We find that our employed generic conditions *background, position, shape, action* consistently yield more performance gains (0.51%∼1.38%) than *pattern, anchor, surface, habits*. This experiment primarily verifies that semantic and generalization disparities also exist among LLM-generated conditions. Due to space constraints and research focus considerations, we refrain from further exploring subsequent cleaning of these conditions and simply rely exclusively on the optimal condition identified in our ablation studies for all experiments.

Ablation study of condition order. To investigate this factor, we take *background, position, shape, action* and reorder them in the prompts. The results are shown in Table 8. We can see that changing the order of conditions does not substantially affect the model's performance, but placing *position* at the beginning seems to cause a minor degradation. We suggest that this may result from CLIP focusing more on earlier text tokens than later ones (an inherent bias of CLIP proposed by [17]), and *position* being comparatively harder to perceive than others.

E. Limitations and Broader Impacts

Although employing conditions to intervene in MLR and learn non-spurious correlations is inspiring, not all conditions can be equally perceived by the VLM. Our ablation study in Sec. D verifies this: some salient visual conditions like pose, color, and size contribute prominently to the overall performance, while others like symmetry or habits are relatively hard to perceive. This also explains why our model's performance degrades when we ask the LLM to summarize an excessive number of conditions and utilize them in the prompts, as a significant portion of these conditions are redundant and ambiguous. Second, simply leveraging a few learnable tokens to learn condition content may be insufficient in modeling capacity (however, expanding the learnable modules may also dramatically increase complexity). We hope future endeavors will focus on generating more robust and less biased conditions, achieving a better trade-off between efficiency and performance.

From another perspective, this paper treats Multi-Label Recognition (MLR) only as a classification task; other MLR tasks like multi-label object detection and semantic segmentation remain unexplored. Whether these tasks would encounter similar performance degradation when combined with FL should be carefully considered.

References

[1] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009. 1, 2

[2] Tianyu Cui, Hongxia Li, Jingya Wang, and Ye Shi. Harmonizing generalization and personalization in federated prompt learning. *arXiv preprint arXiv:2405.09771*, 2024. 2

[3] Zixuan Ding, Ao Wang, Hui Chen, Qiang Zhang, Pengzhang Liu, Yongjun Bao, Weipeng Yan, and Jungong Han. Exploring structured semantic prior for multi label recognition with incomplete labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3398–3407, 2023. 2, 3

[4] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 2

[5] Ping Hu, Ximeng Sun, Stan Sclaroff, and Kate Saenko. Dualcoop++: Fast and effective adaptation to multi-label recognition with limited annotations. *IEEE Transactions on Pat-*

Table 9. Lists of conditions under varied requirement number in LLM-queries.

Number	LLM-generated conditions	mAP (%)
1	[“context”]	87.14
2	[“context”, “layout”]	88.77
3	[“background”, “position”, “action”]	89.61
4	[“background”, “position”, “shape”, “action”]	90.10
5	[“color”, “texture”, “shape”, “contrast”, “action”]	90.06
6	[“background”, “layout”, “context”, “proximity”, “activity”, “setting”]	89.98
8	[“Background”, “Lightness”, “Color”, “Texture”, “Shape”, “Size”, “Position”, “Action”]	90.12
10	[“Background”, “Lightness”, “Texture”, “Shape”, “Color”, “Size”, “Action”, “Position”, “Occlusion”, “Perspective”]	90.07
12	[“background”, “lightness”, “color”, “texture”, “shape”, “size”, “Action”, “perspective”, “contrast”, “position”, “shadow”, “context”]	89.13
14	[“background”, “lightness”, “color”, “texture”, “shape”, “size”, “Action”, “pattern”, “contrast”, “position”, “depth”, “orientation”, “symmetry”, “context”]	89.51
16	[“background”, “lightness”, “contrast”, “texture”, “shape”, “size”, “color”, “pattern”, “orientation”, “position”, “Action”, “depth”, “symmetry”, “sharpness”, “transparency”, “reflectance”]	89.48
18	[“color”, “texture”, “shape”, “size”, “pattern”, “contrast”, “symmetry”, “depth”, “perspective”, “Action”, “transparency”, “reflection”, “shadow”, “focus”, “alignment”, “density”, “complexity”, “clarity”]	89.22
20	[“background”, “lightness”, “color”, “texture”, “shape”, “size”, “Action”, “direction”, “position”, “depth”, “perspective”, “contrast”, “pattern”, “symmetry”, “occlusion”, “context”, “material”, “reflection”, “transparency”, “shadow”]	89.24

tern Analysis and Machine Intelligence, 46(5):3450–3462, 2023. 1

[6] Y. Hua, L. Mou, P. Jin, and X. X. Zhu. Multiscene: A large-scale dataset and benchmark for multi-scene recognition in single aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, in press. 2

[7] 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, 3

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2

[9] Xiaoman Qi, Panpan Zhu, Yuebin Wang, Liqiang Zhang, Junhuan Peng, Mengfan Wu, Jialong Chen, Xudong Zhao, Ning Zang, and P Takis Mathiopoulos. Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:337–350, 2020. 2

[10] Chen Qiu, Xingyu Li, Chaithanya Kumar Mummadi, Madan Ravi Ganesh, Zhenzhen Li, Lu Peng, and Wan-Yi Lin. Federated text-driven prompt generation for vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3

[11] Samyak Rawlekar, Shubhang Bhatnagar, and Narendra Ahuja. Positivecoop: Rethinking prompting strategies for multi-label recognition with partial annotations. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5863–5872. IEEE, 2025. 2, 3

[12] Changlong Shi, He Zhao, Bingjie Zhang, Mingyuan Zhou, Dandan Guo, and Yi Chang. Fedawa: Adaptive optimization of aggregation weights in federated learning using client vectors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30651–30660, 2025. 2, 3

[13] Mainak Singha, Subhankar Roy, Sarthak Mehrotra, Ankit Jha, Moloud Abdar, Biplab Banerjee, and Elisa Ricci. Fedmvp: Federated multi-modal visual prompt tuning for vision-language models. *arXiv preprint arXiv:2504.20860*, 2025. 2, 3

[14] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *Advances in Neural Information Processing Systems*, 35:30569–30582, 2022. 1, 2, 3

[15] Hao Tan, Zichang Tan, Jun Li, Ajian Liu, Jun Wan, and Zhen Lei. Recover and match: Open-vocabulary multi-label recognition through knowledge-constrained optimal transport. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4650–4660, 2025. 1, 2, 3

[16] Hantao Yao, Rui Zhang, and Changsheng Xu. Tcp: 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. 2, 3

[17] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pages 310–325. Springer, 2024. 4