FedMPT: Federated Multi-Label Prompt Tuning of Vision-Language Models

Anonymous CVPR submission

Abstract

Multi-Label Recognition (MLR) based on Vision-Language Models (VLMs) aims to leverage their pre-trained knowledge to better adapt complex recognition scenarios, thereby enhancing model robustness. However, for realistic decentralized applications requiring federated learning, adapting VLMs to each client that possesses private and heterogeneous data can cause the model to overfit spurious label correlations, consequently triggering irrelevant categories when encountering new samples. To tackle this problem, we reconsider the federated learning for MLR with a causal model, in which we adopt a front-door adjustment and decouple the MLR modeling process by intermediate variables that magnify the oracle label co-occurrence. Guided by our analysis, we propose our FedMPT, the first method specifically designed for federated MLR. The core idea of FedMPT is to leverage generalizable conditions to steer federated MLR to mitigate erroneous label activations. To achieve this, FedMPT introduces an Large Language Model (LLM)-driven pipeline to decipher the underlying conditions that govern label dependencies. Furthermore, we introduce an optimal transport between the condition-enriched prompts and the image patches to uncover multiple region-level semantics. Finally, we generate synergistic predictions from different conditions with a crafted gating mechanism. Experiments on multiple benchmark datasets show that our proposed approach achieves competitive results and outperforms SOTA methods under varied settings. Code is available in the ***Supplementary Material****.*

1. Introduction

Multi-Label Recognition (MLR) aims to identify all possible labels in a single image. Owing to its alignment with real-world requirements, MLR has found wide application [4, 5, 21, 63]. Early methods primarily focused on modeling inter-label co-occurrences [9, 10, 15, 55], refining the models' attention on local regions [17, 60] or balancing the positive-negative gradients [46]. Recently, emerging efforts incorporate prompting pretrained Vision-Language Models (VLMs) [23, 29, 30, 33, 43] for MLR owing to their remarkable zero-shot generalization abilities learned from web-scale image-text pairs. For instance, DualCoOp [51] and PosCoOp [44] adapt the prompt learning to MLR by learning two collaborative prompts for each class; ML-VPT [37] introduces distinct prompts for correlative and distinctive classes; SPARC [39] and CCD [26] unravel the inherent bias of VLMs to MLR as uneven score distribution across labels, then mitigate this bias for enhanced zero-shot learning and knowledge distillation respectively.

Although most MLR methods are designed for centralized settings where the model has full access to the dataset space, real-world applications often necessitate a decentralized architecture [12, 38, 48, 50, 65] (i.e., Federated Learning, FL) where each client only possesses a heterogeneous

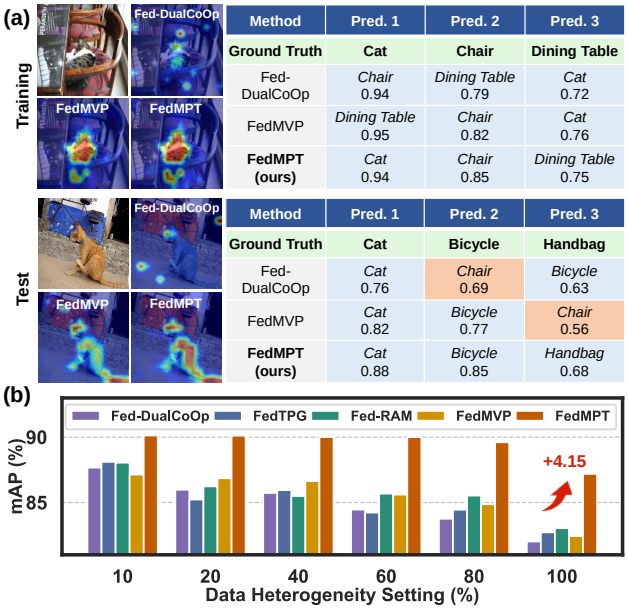

Figure 1. **(a):** Comparison of class-activation map for "`Cat`" and top-3 predictions on the training image **(a, upper)** and test image **(a, lower)**. Existing SOTAs are prone to overfitting spurious correlation (i.e., `cat-chair`) and diverting attentions under FL, while our FedMPT effectively alleviates these issues. **(b):** As data heterogeneity increases, existing SOTAs show significant degradation, while our FedMPT demonstrates substantial robustness.

and private portion of the data. The long-standing research focuses of FL (with VLMs) lies on modeling the commonality and specificity of client distributions, for example, FedCoOp [18], PromptFL [18] and FedTPG [42] introduce shared prompts with FedAvg [38] to learn generic knowledge across clients; FedPGP [12] and FedOTP [28] introduces the local-global prompt collaboration for balancing generic and customized modeling. ***Notably, to the best of our knowledge, all existing VLM-based FL methods are built for single-label recognition and consistently overlook the practical challenges of MLR***.

Diving into the integration of MLR and FL, we present a critical two-fold ***dilemma***: first, if we directly train MLR SOTAs [44, 48, 52] on each client and aggregate their weights via FedAvg, the global model would learn excessively spurious [37] label correlations and show severe performance degradation under increasing data heterogeneity [47]. Second, conventional FL methods are ill-suited to MLR, as they fail to capture the inherent inter-dependencies between labels, similarly resulting in correlation overfitting and incomplete retrieval. Figure 1.a illustrates an example: the aggregated global model of existing SOTAs (DualCoOp [51] and FedMVP [48]) overfits to the `cat-chair` correlation, spuriously boosting the `chair` score upon seeing a `cat` in inference, meanwhile diverting their prediction confidence to ground-truth labels. Figure 1.b summarizes mAP of different methods under varied data heterogeneity across clients (induced via clustering). We observe that existing SOTAs unanimously show sharp degradation despite their strong performance in near-IID settings (i.e., 10%).

To further understand MLR under FL, we reconsider it with a Structural Causal Modeling (SCM) in Section 3.2. Our key findings are that semantic variables learned from pre-training and control the content of images and labels, can be naturally divided into generic and client-specific variants. Overfitting the latter would lead to degraded generalization. Then, from the perspective of front-door adjustment, our objective is to identify an intermediate variable that maximizes the oracle label correlations.

Guided by our analysis, we propose our FedMPT, a novel condition-driven framework specifically designed for MLR under FL. FedMPT is built on a foundational idea: to leverage multiple, complementary conditions to intervene the MLR tasks. Concretely, we first devise an LLM-driven pipeline to generate generic abstract condition templates, which are incorporated into prompts for soft prompt learning; These condition prompts are then aligned with relevant image regions via optimal transport to produce multiple diverse, condition-specific predictions. Finally, inspired by the expert routing mechanism in LLMs [3, 8], we introduce an adaptive gating mechanism to automatically adjust condition contributions in each client. We incorporate the Asymmetric Loss (ASL) as our training objective. In one communication round, all clients share their learnable parameters with the server, which aggregates them via FedAvg to form a unified global model. Comprehensive evaluations on three MLR benchmarks across various federated settings demonstrate that FedMPT substantially outperforms existing SOTA methods and exhibits remarkable robustness. Additional analyses validate the efficiency of our approach. Our contribution can be summarized as:

- We identify and formalize the novel problem of Multi-Label Recognition (MLR) under realistic federated scenarios and unveil the vulnerability of existing methods.
- From a causal perspective, we attribute the intricacy of MLR under FL as the overfitting to local distributions and label correlations. Guided by our analysis, we propose FedMPT, which leverages multiple conditions to synergistically learn generic semantics across clients.
- Extensive experiments on three benchmarks show that FedMPT achieves state-of-the-art performance and exhibits remarkable robustness under various federated settings. Further ablations to highlight its efficiency.

2. Related Work

Multi-Label Recognition (MLR). Multi-Label Recognition (MLR) demands the precise identification of all relevant labels in an image and naturally has broad real-world applications. Traditional MLR methods approaches have primarily focused on modeling class specifications and correlations; One line of work [7, 10, 54] incorporate text embedding graphs of labels to model the similarity of classes; Another line [22, 36, 40, 45] dives into the local regions and discover the class-specific visual cues from each crop; More recently, the advancement of Vision-Language Models (VLMs) like CLIP has inspired a new direction for MLR: For instance, CDUL [1] devises an unsupervised framework where global and local knowledge are fused to generate pseudo labels for unlabeled samples; Concurrently, SPARC [39] and CCD [26] identify the inherent class bias in VLMs and explicitly mitigates this bias to achieve superior performance. Another notable progress lies in incorporating prompt learning to VLM-MLRs, for example, DualCoOp [51] and DualCoOp++ [19] introduce two prompts to model the object existence/non-existence in each image patch; RAM [52] introduces a local knowledge guided aggregation scheme for open-vocabulary MLR; PosCoOp [44] enhances DualCoOp with an unconditioned prompt for object absence modeling.

Federated Learning (FL) with VLMs. Federated Learning (FL) [38, 48, 50, 65] has emerged as a pivotal paradigm for enabling decentralized and privacy-preserving training on heterogeneous data. The application of FL to VLMs has seen significant evolution. Initial methods introduce prompt learning on each client with FedAvg to aggregate prompt weights [18]; Subsequent research incorporates granular

learnable modules like adapters [35] or prompt generators [13, 42, 48]. For example, FedMVP [48] generates visual embeddings with image tokens and LLM-attribute embeddings through a specialized cross-modal PromptFormer; Some other methods endeavor to harmonize the local and global knowledge [12, 28], for example, FedOTP [28] restrains the contribution of local/global prompts via margin-adapted optimal transport. Concurrently, FL is being explored in various specialized domains, including continual learning [61, 64], test-time adaptation [2, 24], autonomous driving [27], and interpretability [34].

3. Preliminaries and Problem Analysis

3.1. Preliminaries

Multi-Label Recognition (MLR) with VLMs. We first summarize a baseline [44, 51] for MLR with VLMs built on prompt learning: given a typical VLM (CLIP) that employs dual encoders for processing multi-modal information, let $\mathcal{E}_v$ and $\mathcal{E}_t$ denote the image and text encoder respectively; Given the input $(\boldsymbol{x}, \boldsymbol{y})$ where label $\boldsymbol{y} \in \mathbb{R}^C$ (C is the number of classes), CLIP encodes $\boldsymbol{x}$ into M-length visual embeddings $\boldsymbol{v}^0$ with $\mathcal{E}_v$; for the text modality, we fill all classnames into learnable templates (initialized as *A photo of a [CLASS]*), yielding prompt $\boldsymbol{p}$; $\boldsymbol{p}$ is encoded into text embeddings $\boldsymbol{t}$ with $\mathcal{E}_t$. The operations at layer i of $\mathcal{E}_v$, $\mathcal{E}_t$ are:

$$\begin{aligned}[\texttt{cls}^i, \boldsymbol{v}^i] &= \mathcal{E}_v^i([\texttt{cls}^{i-1}, \boldsymbol{v}^{i-1}]), \boldsymbol{v}^0 = Emb(\boldsymbol{x}) \\ [\texttt{bot}^i, \boldsymbol{p}^i, \texttt{eot}^i] &= \mathcal{E}_t^i([\texttt{bot}^{i-1}, \boldsymbol{p}^{i-1}, \texttt{eot}^{i-1}]),\end{aligned} \tag{1}$$

where $[\cdot, \cdot]$ means concatenation, $\texttt{cls}, \texttt{bot}, \texttt{eot}$ represent the cls (class), bot (begin-of-text) and eot (end-of-text) tokens. Emb is the patch embedding layer, $\boldsymbol{v}^i$ is the patch embeddings of layer i. The output projection P_t is applied to $\texttt{eot}^L$ to generate the text embedding, i.e., $\boldsymbol{f}_t = P_t(\texttt{eot}^L)$. For the visual modality, instead of using the global representation $\texttt{cls}^L$, this baseline projects $\boldsymbol{v}^L$ into the final patch-level output embeddings $\boldsymbol{f}_v(\boldsymbol{v})$, i.e., $\boldsymbol{f}_v(\boldsymbol{v}) = P_v(\boldsymbol{v}^L)$. The final prediction is calculated by selectively aggregating the predictions over patches (similarity between $\boldsymbol{f}_v(\boldsymbol{v})$ and $\boldsymbol{f}_t$) based on their softmax-normalized weights:

$$\mathbb{P}(y_c|\boldsymbol{x}) = \sum_m \frac{\exp(s_{m,c}/\tau)}{\sum_{c'} \exp(s_{m,c'}/\tau)} \cdot s_{m,c}, \tag{2}$$

where $\texttt{sim}(\cdot, \cdot)$ represents the cosine similarity in default, $s_{m,c} = \texttt{sim}(\boldsymbol{f}_v(\boldsymbol{v}_m), \boldsymbol{f}_{t,c})$ is the cosine similarity between patch m and class c. τ is the temperature. The final logits are optimized with the Asymmetric Loss (ASL) [46] to handle optimization imbalance of positive and negative classes:

$$\mathcal{L}_{asl} = (1 - \mathbb{P})^{\gamma_+} \boldsymbol{y} \log(\mathbb{P}) + (\mathbb{P}^c)^{\gamma_-} (1 - \boldsymbol{y}) \log(1 - \mathbb{P}^c) \tag{3}$$

where $\mathbb{P}^c = \max(\mathbb{P} - c, 0)$ is for truncating negative predictions, which is controlled by the hard threshold c. We set the hyper-parameters as $\gamma_- \geq \gamma_+$, so that ASL would better down-weight the contribution of easy negative samples.

Federated Learning (FL). Following previous studies [42, 48], we consider a standard FL system comprising K remote client models $\{\boldsymbol{\rho}_k\}_{k=1}^K$ running optimization on their local data D_k, as well as a server $\mathcal{G}$ for coordination by aggregating and broadcasting parameters. We follow a non-IID federated setup where local data for different client are heterogeneous; *to achieve this under MLR settings, we cluster the dataset based on the image features extracted from zero-shot ViT/B-16, then assign each cluster to one client.* The objective is to learn an optimal global model $\boldsymbol{\rho}$ aggregated from clients with the minimum risk $\mathcal{L}$, undergoing ϕ communication rounds with a client participation rate of ϵ:

$$\mathcal{L} = \min_{\boldsymbol{\rho}} \sum_{k=1}^K p_k \mathcal{L}_k(\boldsymbol{\rho}, \mathcal{D}_k; \mathcal{G}[\phi; \epsilon]), \tag{4}$$

where p_k represents the weight of k-th client and set to $|\mathcal{D}_k| / \sum_{\mathcal{D}_l \in \mathcal{D}_k} |\mathcal{D}_l|$, where $|\mathcal{D}_k|$ is the size of $\mathcal{D}_k$.

3.2. Problem Analysis

This subsection formalizes the problem of MLR under FL from a causal perspective. Our proposed Structural Causal Model (SCM) is depicted in Figure 2, where nodes represent variables during pre-training or fine-tuning, and edges denote causal relationships. Unobservable and observable variables are highlighted in red and gray, respectively.

Concretely, $\mathcal{D}_o$ means the pre-training data, which determines the semantic factors $\mathcal{F}$ which controls the semantics of input space $\mathcal{D}$ and output space $\mathcal{Y}$. As shown in Figure 2.a, under the federated learning scenario, $\mathcal{F}$ can be divided into generic factors $\mathcal{F}_g$ which capture transferable knowledge across clients and are the target of finetuning, and $\mathcal{F}_s$, which encapsulate client-specific semantics that may induce overfitting to local spurious correlations. Notably, $\mathcal{F}_s$ is also influenced by manual factors $\mathcal{M}$ like data partitioning policies. The image content is a mixture of both, whereas the labels can only be derived from the generic factors $\mathcal{F}_g$.

Our objective is to maximize the influence of $\mathcal{F}_g$ while minimizing that of $\mathcal{F}_s$ during training. This enables an unbiased estimation of the causal effect $\mathcal{X} \to \mathcal{Y}$. However, the insufficiency of local training data and its dramatic gap with inference data makes our modeling biased and capturing a collection $\mathcal{F}_{g,s}$ of both $\mathcal{F}_g$ and $\mathcal{F}_s$ (Figure 2.b), resulting in a backdoor-path of $\mathcal{D} \leftarrow \mathcal{F}_{g,s} \to \mathcal{Y}$. To tackle this issue, we incorporate a well-known front-door adjustment [56, 62] by introducing an intermediate variable $\mathcal{R}$ (Figure 2.c) that ideally reflects the causality between $\mathcal{X}$ and $\mathcal{Y}$. We can then identify $P(\mathcal{Y}|do(\mathcal{X}))$ with front-door adjustment:

$$P(\mathcal{Y}|do(\mathcal{X})) = \mathbb{E}_{P(r|x)} \mathbb{E}_{P(x')} P(Y|r, x') \tag{5}$$

where x, r denote specific values of $\mathcal{X}$ and $\mathcal{R}$. The core challenge thus reduces to constructing r that accurately captures the oracle causal mechanism $\mathcal{X} \to \mathcal{Y}$. In the next section, we'll introduce our FedMPT, which incorporates condition-guided learning and gating to meet our analysis.

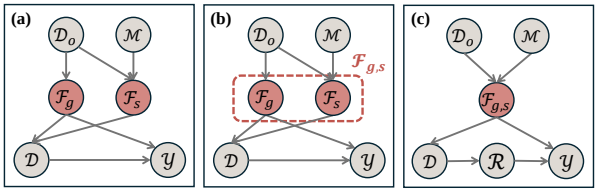

Figure 2. Structural Causal Model (SCM) for MLR under FL.

4. Methodology

This section introduces our proposed FedMPT, comprising condition prompt generation (§4.1), condition-guided optimal transport (§4.2), condition gating (§4.3), and the federated communication process (§4.4).

4.1. Condition Prompt Generation

How can we maximize the adjustment of variable r of SCM in §3.2? Since directly learning from the datasets leads to spurious correlations and degraded generalization, we propose to intervene MLR with certain ***conditions*** that approximate the oracle causalities and label correlations: Recalling the instance of Figure 1.a, we tend to accept the cat-chair concurrent predictions under conditions of *"indoor scene"*, *"wooden textures"*, and *"lying actions"*; Thus, the model will reduce `cat`'s weight when faced with a (table, beach) image, where some ***conditions*** are not satisfied.

Pursuing this idea, our goal is to generate a set of generic, broad, and fine-grained conditions that can be shared across all clients. Our strategy is to fix some abstract conditions and leave the specific and contextualized contents learnable. To generate the abstract conditions, we employ an LLM-driven pipeline (Figure 3) with Chain-of-Thought (CoT) following [31]. Concretely, we first prompt the LLM to generate as many descriptions as possible for each possible combination of dataset categories; thanks to the rich knowledge embedded in LLMs, we can acquire the characteristics and existence conditions of various label combinations at this stage. Then, we prompt LLM to summarize N distinct abstract conditions which would exactly encapsulate label correlations; finally, we obtain several abstract conditions like *"spatial layout"*, *"object pose"*, *"background"*, etc.

To integrate these abstract conditions into the learning process, we populate them into the "[COND]" slot of the following template, yielding prompts $\boldsymbol{p}^\dagger = \{\boldsymbol{p}_1^\dagger, ..., \boldsymbol{p}_C^\dagger\}$:

$$[L_1]\cdots[L_{\beta_{cond}}]\ \text{[COND]}\ [L_1]\cdots[L_{\beta_{cls}}]\ \text{[CLASS]},$$

where $[L_\cdot]$ means the learnable tokens, β_{cond}, β_{cls} control the number of condition-level and class-level tokens respectively. Critically, the former ones are specified for each condition, while the latter ones are shared by all classes. $\boldsymbol{p}^\dagger$ is maintained in the server and distributed to all clients in the communication phase. We denote $\boldsymbol{f}_t(\boldsymbol{p}^\dagger)$ as the output text embeddings of $\boldsymbol{p}^\dagger$ processed by the text encoder. Concrete conditions and more discussions are in **Sup. Mat. D**.

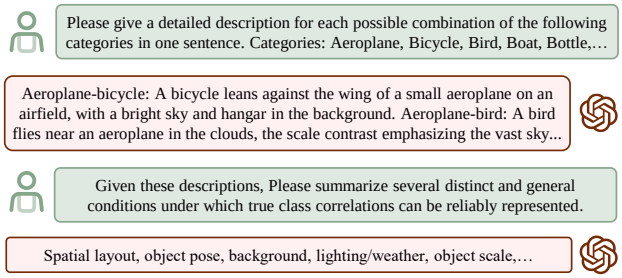

Figure 3. Our proposed LLM-based condition generation pipeline.

4.2. Condition-guided Optimal Transport

To better align the generated condition prompts with region-level fine-grained visual semantics, we devise an Optimal-Transport (OT) between output patch embeddings (denoted as $\boldsymbol{f}_v(\boldsymbol{v})$ in the above) and the condition prompts received from the server. Specifically, we first introduce M adapters $\{\mathcal{A}_m\}_{m=1}^M$ (corresponding to attributes) over $\boldsymbol{f}_v(\boldsymbol{v})$ to generate new visual latent spaces, where each adapter is a LoRA-like [53, 58] architecture for efficiency:

$$\boldsymbol{f}_{v,n}^\dagger(\boldsymbol{v}) = \mathcal{A}_n(\boldsymbol{f}_v(\boldsymbol{v})) = W_\uparrow(W_\downarrow(\boldsymbol{f}_v(\boldsymbol{v}))), \tag{6}$$

where $W_\downarrow \in \mathbb{R}^{D\times D_s}$,$W_\uparrow \in \mathbb{R}^{D_s\times D}$, D/D_s is the output / down-projected latent dimension. The OT aims to find an optimal plan $\mathcal{P}^* \in \mathbb{R}^{M\times N\times C}$ that minimizes the distance between distributions, i.e., $\mathcal{P}^* = \texttt{OT}(\mathcal{C}; \boldsymbol{a}, \boldsymbol{b})$, where $\mathcal{C} \in \mathbb{R}^{M\times N\times C}$ represents the cost-matrix, $\boldsymbol{a} \in \mathbb{R}^N$, $\boldsymbol{b} \in \mathbb{R}^M$ are constrained marginal distributions. $\mathcal{C}$ is calculated with:

$$\mathcal{C}_{m,n} = 1 - \frac{\exp(\texttt{sim}(\boldsymbol{f}_{v,n}^\dagger(\boldsymbol{v}_m), \boldsymbol{f}_t(\boldsymbol{p}_n^\dagger))/\tau)}{\sum_m \exp(\texttt{sim}(\boldsymbol{f}_{v,n}^\dagger(\boldsymbol{v}_m), \boldsymbol{f}_t(\boldsymbol{p}_n^\dagger))/\tau)}. \tag{7}$$

We also denote $\mathcal{S} = 1 - \mathcal{C}$, i.e., the original region-text similarity. We keep $\boldsymbol{b}$ as the uniform distribution so that all categories yield equal change to be detected in the image. For $\boldsymbol{a}$, inspired by recent finding [6] indicating that regions yield different contributions of semantics, we set $\boldsymbol{a}$ as the semantic importance of each patch, calculated by:

$$a_{m,n} = \frac{\exp(H(\texttt{sim}(\boldsymbol{f}_{v,n}^\dagger(\boldsymbol{v}_m), \boldsymbol{f}_t(\boldsymbol{p}_n^\dagger)))/\tau)}{\sum_m \exp(H(\texttt{sim}(\boldsymbol{f}_{v,n}^\dagger(\boldsymbol{v}_m), \boldsymbol{f}_t(\boldsymbol{p}_n^\dagger)))/\tau)}, \tag{8}$$

where H denotes the self-entropy. To calculate OT more efficiently, we introduce entropy relaxation to approximate the results with the Sinkhorn [49] algorithm, formulated as:

$$\mathcal{P} = \texttt{diag}(\mathcal{U})\mathcal{M}\texttt{diag}(\mathcal{V}),\ \mathcal{U} = \{u_m\}_{m=1}^M, \mathcal{V} = \{v_n\}_{n=1}^N, \tag{9}$$

where $\mathcal{U}$, $\mathcal{V}$ are updated with the following recurrent form:

$$u_m \leftarrow \frac{\boldsymbol{a}}{\sum_n \mathcal{M}_{m,n} v_n}, v_n \leftarrow \frac{\boldsymbol{b}}{\sum_n \mathcal{M}_{m,n} u_n}, \mathcal{M} = \exp(\frac{-\mathcal{C}}{\lambda}) \tag{10}$$

With the OT applied across all classes for each conditioned prompt, we finally obtain N similarity maps between regions and classes. For each client, we calculate the Wasserstein distance $\boldsymbol{\psi} \in \mathbb{R}^{N\times C}$ by class for each condition,

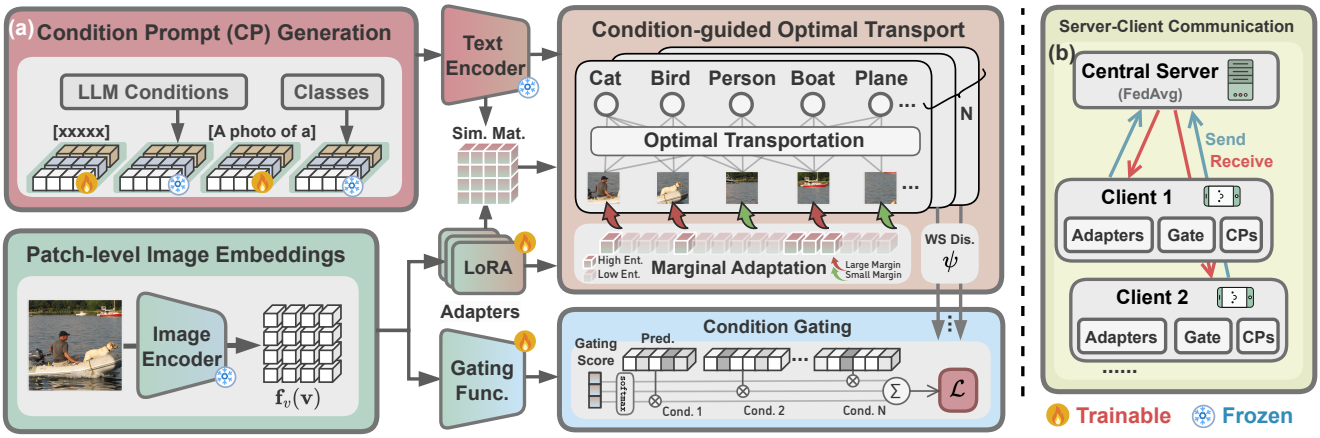

Figure 4. Overview of our proposed FedMPT framework. (a) The LLM-generated conditions are instantiated into Condition Prompts (CPs), which are encoded into text embeddings. For a given image, its visual feature map is aligned with these prompt embeddings via Optimal Transport (OT). The contributions of different conditions are then adaptively calibrated by a gating module. (b) At each communication round, the server aggregates the parameters of CPs, adapters, and gating modules and distributes the updated parameters back.

which reflects the affinity of each class to the visual regions and calculated via $\boldsymbol{\psi}_n = \sum_m \mathcal{P}_{m,n}\mathcal{S}_{m,n}$. Consequently, we treat $\boldsymbol{\psi}$ as conditioned predictions, i.e., $\mathbb{P}_n = \boldsymbol{\psi}_n$.

4.3. Condition Gating

While the conditional prompting and OT matching mitigate overfitting to local spurious correlations in MLR, the relevance of each condition may not always remain the same across clients due to data heterogeneity. To ensure robust generalization, we introduce a gating mechanism that dynamically adapts the influence of each condition. Specifically, inspired by Mixture-of-Experts (MoE) [53, 57] in LLMs, we leverage a router $\omega \in \mathbb{R}^{D\times N}$ to dynamically determine the contribution of different conditions and aggregate their predictions:

$$\omega = \Omega(\boldsymbol{f}_v(\boldsymbol{v})); \quad \mathbb{P}' = \sum\nolimits_n \frac{\exp(\omega_n)}{\sum_{n'}\exp(\omega_{n'})}\mathbb{P}_n, \tag{11}$$

where $\Omega \in \mathbb{R}$ is a similar LoRA module like $\{\mathcal{A}_i\}$.

4.4. Local Training and Federated Average

Local Training. Each client optimizes the local model using their private data via the above asymmetric loss (ASL):

$$\mathcal{L} = (1-\mathbb{P}')^{\gamma_+}\boldsymbol{y}\log(\mathbb{P}') + (\mathbb{P}'_c)^{\gamma_-}(1-\boldsymbol{y})\log(1-\mathbb{P}'_c) \tag{12}$$

Recall that γ_+,γ_- are hyper-parameters to control the contribution of positive/negative regularizations. Ablations of these hyper-parameters are in **Sup. Mat. B.**

Federated Average. In one communication round, the server collects the updated weights of condition prompts $\boldsymbol{p}$, adapters $\{\mathcal{A}_m\}_{m=1}^M$ and gates Ω from clients, where they're aggregated among different clients $\texttt{Cli}_t$ with FedAvg [38]:

$$\{\boldsymbol{p}, \{\mathcal{A}_m\}_{m=1}^M, \Omega\} \leftarrow \frac{1}{K}\sum\nolimits_t \texttt{Cli}_t(\{\boldsymbol{p}, \{\mathcal{A}_m\}_{m=1}^M, \Omega\}) \tag{13}$$

The aggregated weights are sent to clients for the subsequent training. The above steps are repeated for R rounds.

5. Experiments

Benchmarks. We systematically evaluate the effectiveness of FedMPT under the federated MLR setting with three benchmarks: ❶ Heterogeneity Benchmark, which assesses model robustness to varying degrees of data heterogeneity across clients. To achieve this goal, the training dataset is first partitioned into S clusters according to their visual embeddings from ViT/B-16, then each data cluster is assigned to a client. We change S by varying $t(\%)$, the proportion of the class size C. ❷ Federated Part-Annotation Benchmark, which evaluates the models' robustness to insufficient annotations by randomly masking '`Mask`' class annotations in the training set. The heterogeneity setting t is kept as 60% for this benchmark. We employ three datasets for the above benchmarks: VOC2007 [16], COCO2014 [32], and NUS-wide [11]. Furthermore, to assess models' real-world applicability, we adopt ❸ Federated Real-world MLR Benchmark[52], which incorporates two remote sensing datasets: MultiSense [20] and MLRSNet [41]. $t = 60\%$. *We evaluate the performance of the global model on the test sets. (For methods [12] that also maintain local private parameters in clients, we evaluate the performance of each client on the test sets and average them).* Following [37], we report the three most important metrics in all benchmarks: Mean Average Precision (mAP), per-category F1-score (CF1), and overall F1-score (OF1). All reported results are the average of 3 independent runs. Experiments on other datasets, finer settings, or benchmarks (like ZSL [51]) are in **Sup. Mat. A.**

We select the following competitive baselines spanning MLR, FL and Prompt Learning for comprehensive comparisons: DualCoOp [51], SCPNet [14], PosCoOp [44] and RAM [52], which are competitive baselines of MLR with VLMs; MaPLe [25]&TCP [59], which are typical repre-

Table 1. Results on the Heterogeneous Benchmark. We report the mAP, per-category F1 (CF1) and overall F1 (OF1) with the client number varies from 10% to 100% of the class number. The best results are marked with **bold**.

VOC2007

Methods	Venue	$t = 10\%$			$t = 20\%$			$t = 40\%$			$t = 60\%$			$t = 80\%$			$t = 100\%$			Avg		
		mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1
Fed-DualCoOp	NeurIPS'22	87.67	81.32	81.91	85.98	78.84	76.16	85.73	78.33	76.44	84.46	77.13	75.25	83.76	75.44	73.89	82.02	75.00	74.73	84.94	77.68	76.40
Fed-SCPNet	CVPR'23	83.37	78.74	77.86	81.15	74.68	76.05	80.03	74.05	74.09	79.89	66.04	73.45	76.51	74.29	71.13	76.48	74.06	70.64	79.57	73.64	73.87
Fed-MaPLe	CVPR'23	84.22	78.00	82.36	87.87	81.27	80.70	84.36	76.35	75.82	81.82	73.73	73.55	80.71	70.65	72.04	76.08	68.13	70.89	82.51	74.69	75.89
FedPGP	ICML'24	85.47	77.97	79.81	84.51	76.76	77.66	84.47	75.99	73.97	79.89	72.72	67.48	76.60	68.71	66.52	78.16	69.30	68.35	81.52	73.58	72.30
Fed-TCP	CVPR'24	81.56	76.02	79.23	81.15	72.41	79.50	80.65	78.93	80.44	77.32	78.38	78.89	76.50	76.64	73.50	76.42	71.52	76.48	78.93	75.65	78.01
FedTPG	ICLR'24	88.11	81.13	78.05	85.23	81.26	77.73	85.95	79.05	78.54	84.23	77.37	78.68	84.45	77.29	76.13	82.72	76.67	75.87	85.12	78.80	77.50
Fed-PosCoOp	WACV'25	87.42	80.67	79.51	84.71	82.77	77.86	82.53	80.90	76.97	81.65	80.31	76.43	80.71	80.17	75.79	79.90	78.26	75.08	82.82	80.51	76.94
FedAWA	CVPR'25	84.75	78.57	80.41	83.17	76.06	78.35	81.77	75.11	79.99	79.48	72.26	79.24	78.81	70.49	77.68	78.22	70.73	70.15	81.03	73.87	77.64
Fed-RAM	CVPR'25	88.05	81.50	81.98	86.24	81.27	80.70	85.50	79.74	80.01	85.68	79.24	78.90	85.53	79.56	72.84	83.04	77.00	68.99	85.67	79.72	77.24
FedMVP	ICCV'25	87.14	80.24	82.30	86.86	79.32	80.22	86.64	79.00	79.93	85.61	79.37	80.65	84.87	79.23	75.60	82.43	79.56	77.92	85.59	79.45	79.44
FedMPT	Ours	**90.13**	**84.33**	**86.82**	**90.12**	**84.91**	**83.23**	**90.01**	**84.42**	**84.31**	**90.00**	**83.88**	**84.56**	**89.61**	**83.72**	**80.36**	**87.19**	**81.31**	**82.43**	**89.51**	**83.76**	**83.62**
Δ Prev. best	\	**+2.02**	**+2.83**	**+4.46**	**+2.25**	**+2.14**	**+2.53**	**+3.37**	**+3.52**	**+3.87**	**+4.32**	**+3.57**	**+3.91**	**+4.08**	**+3.55**	**+2.68**	**+4.15**	**+1.75**	**+4.51**	**+3.84**	**+3.25**	**+4.18**

COCO2014

Methods	Venue	$t = 10\%$			$t = 20\%$			$t = 40\%$			$t = 60\%$			$t = 80\%$			$t = 100\%$			Avg		
		mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1
Fed-DualCoOp	NeurIPS'22	61.88	58.34	62.15	58.39	54.11	59.82	55.95	51.99	57.17	53.80	49.55	58.33	52.28	48.34	57.86	51.46	48.57	56.95	55.63	51.82	58.71
Fed-SCPNet	CVPR'23	60.96	57.96	59.70	57.68	52.91	57.45	52.79	53.09	53.50	50.10	50.73	55.36	48.40	48.84	45.61	46.67	42.25	47.75	52.77	50.96	53.23
Fed-MaPLe	CVPR'23	64.83	61.28	64.45	62.98	58.40	62.16	61.80	58.25	57.43	60.81	57.83	55.89	55.83	51.76	52.54	47.24	49.97	50.13	58.92	56.25	57.10
FedPGP	ICML'24	63.10	61.09	65.36	59.86	58.35	64.78	58.21	57.52	62.16	58.18	55.79	60.87	56.52	52.36	60.85	54.53	53.19	53.54	58.40	56.38	61.26
Fed-TCP	CVPR'24	62.66	61.50	67.16	62.79	57.46	62.69	61.77	53.16	60.13	60.07	52.46	57.72	59.99	50.93	55.78	57.22	47.75	53.56	60.75	53.88	59.51
FedTPG	ICLR'24	63.91	59.39	64.70	61.34	56.05	55.72	60.95	55.47	51.31	60.54	54.52	54.18	54.53	51.63	53.92	50.88	53.18	52.97	58.69	55.04	55.47
Fed-PosCoOp	WACV'25	63.28	58.69	62.57	59.94	55.71	60.91	55.88	50.41	60.13	53.23	49.58	59.37	50.95	50.03	58.04	50.02	48.06	56.09	55.55	52.08	59.52
FedAWA	CVPR'25	65.58	60.72	61.20	62.09	59.73	60.50	61.47	57.58	59.14	60.14	56.67	59.46	59.01	52.69	58.77	57.57	49.65	57.73	60.98	56.17	59.47
Fed-RAM	CVPR'25	66.06	60.98	62.89	63.08	57.97	64.45	59.97	57.97	64.45	60.85	56.95	62.73	60.88	53.00	56.25	55.64	50.37	55.04	61.08	56.21	60.97
FedMVP	ICCV'25	65.76	58.06	64.01	63.19	56.75	65.48	62.26	57.20	63.51	60.60	53.42	62.66	59.56	51.84	59.51	58.48	50.42	55.30	61.64	54.62	61.75
FedMPT	Ours	**67.37**	**62.88**	**67.57**	**65.91**	**60.38**	**67.11**	**65.31**	**59.10**	**66.07**	**63.86**	**58.46**	**64.32**	**63.09**	**57.19**	**63.68**	**62.37**	**56.96**	**62.80**	**64.65**	**59.16**	**65.26**
Δ Prev. best	\	**+1.31**	**+1.38**	**+0.41**	**+2.72**	**+0.65**	**+1.63**	**+3.05**	**+0.85**	**+1.62**	**+3.01**	**+0.63**	**+1.59**	**+2.21**	**+4.19**	**+2.83**	**+3.89**	**+3.77**	**+5.07**	**+3.01**	**+2.78**	**+3.51**

NUS-Wide

Methods	Venue	$t = 10\%$			$t = 20\%$			$t = 40\%$			$t = 60\%$			$t = 80\%$			$t = 100\%$			Avg		
		mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1
Fed-DualCoOp	NeurIPS'22	51.02	45.89	69.00	49.97	44.36	65.30	46.25	41.62	67.02	42.69	37.11	66.10	40.65	35.69	60.28	39.40	34.54	59.04	45.00	39.87	64.46
Fed-SCPNet	CVPR'23	49.56	42.11	65.63	47.61	39.98	59.46	43.31	40.92	62.04	39.28	36.51	60.45	38.61	39.98	58.46	38.49	39.02	53.23	42.81	39.75	59.88
Fed-MaPLe	CVPR'23	54.78	46.69	72.32	53.68	42.13	71.12	51.34	40.69	73.32	51.62	43.40	72.53	50.98	43.68	72.70	51.20	40.99	70.59	52.27	42.93	72.10
FedPGP	ICML'24	55.92	46.48	75.91	53.41	43.27	75.28	53.78	43.07	72.56	49.64	44.82	72.79	48.45	42.58	72.47	43.20	42.72	71.73	50.73	43.82	73.46
Fed-TCP	CVPR'24	54.46	48.23	70.42	51.93	46.57	70.20	50.91	45.65	69.96	48.28	42.47	65.44	47.20	41.75	66.80	48.89	40.50	63.17	50.28	44.20	67.67
FedTPG	ICLR'24	53.01	50.69	73.32	52.78	47.66	72.50	50.00	46.72	69.52	50.40	44.21	70.47	49.92	43.89	67.10	49.46	42.55	65.46	50.93	45.95	69.73
Fed-PosCoOp	WACV'25	52.88	47.96	70.38	50.64	45.89	70.20	48.77	42.69	68.66	42.39	39.15	68.04	40.62	35.12	64.33	40.08	32.76	60.11	45.90	40.60	66.95
FedAWA	CVPR'25	54.67	47.61	74.51	53.91	45.52	70.33	50.40	43.74	71.60	52.87	43.22	71.51	52.75	43.02	66.89	50.71	41.13	66.53	52.55	44.04	70.23
Fed-RAM	CVPR'25	56.35	48.89	76.07	54.05	47.89	76.40	53.51	45.69	74.33	52.25	40.69	74.38	52.61	42.12	72.98	51.20	40.69	70.35	53.33	44.33	74.09
FedMVP	ICCV'25	53.20	48.37	77.80	53.32	46.39	77.10	51.90	45.21	75.14	52.98	43.43	75.02	51.33	42.79	74.55	51.08	40.16	72.90	52.30	44.39	75.42
FedMPT	Ours	**58.36**	**51.56**	**78.98**	**57.27**	**49.29**	**77.93**	**57.03**	**48.14**	**76.73**	**56.52**	**46.88**	**76.94**	**56.08**	**45.73**	**76.60**	**54.87**	**45.03**	**76.81**	**56.69**	**47.77**	**77.33**
Δ Prev. best	\	**+2.01**	**+0.87**	**+1.18**	**+3.22**	**+1.40**	**+0.83**	**+3.25**	**+1.42**	**+1.59**	**+3.54**	**+2.06**	**+1.92**	**+3.33**	**+1.84**	**+2.05**	**+3.67**	**+2.31**	**+3.91**	**+3.36**	**+1.82**	**+1.91**

sentations of prompt learning with VLMs; FedPGP [12], FedTPG [42], FedAWA [47] (For the fairness of comparison, we apply FedAWA to the prompt learner of CoOp [66]) and FedMVP [48], which are state-of-the-arts of federated learning with VLMs. Notably, for methods that are not originally designed for federated scenarios, we alter their methodology by training a local model for each client (with their original cross-entropy loss altered to $\mathcal{L}_{asl}$ mentioned in Sec 3.1) and aggregating the weights of their learnable modules with FedAvg. We add a "Fed-" prefix to the names of these methods to highlight our modification. More details of datasets and baselines are in **Sup. Mat. C.**.

Implementation Details. We employ CLIP (ViT-B/16) with both encoders frozen and the SGD optimizer with a maximum learning rate of 0.001. The batch-size is 32. λ is 0.2. τ=4. The length of learnable tokens for conditions and classes β_{cond}, β_{cls} is 4. The training epoch for VOC2007 and Multi-scene is 100; For COCO2014, NUS-Wide, and MLRSNet, it's 200. A communication round is conducted after one epoch by default. *Epoch and round settings are the same for all methods for fairness.* The client number S and participation rate ϵ vary in different experiments.

5.1. Experiment Results

Results of Heterogeneity Benchmark. We report the results in Table 1, where we draw the following key observations: **(a)** directly transferring standard prompt learning methods like TCP and MaPLe yields unsatisfying performance and robustness to heterogeneity changes (a maximal degradation of about 8.14% in mAP), potentially stemming from their neglect of contextual region semantics; **(b)** although applying MLR methods to federated scenario yields competitive performance, they're comparably vulnerable to increased heterogeneity for severe overfitting to local data; **(c)** SOTAs of federated learning like FedTPG and FedMVP, generally achieve top-tier performance (82.72%, 82.43%)

Table 2. Comparison of FedMPT and other methods on the Federated Part-Annotation Benchmark. 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**.

VOC2007																			
Methods	Venue	Mask = 10%			Mask = 30%			Mask = 50%			Mask = 70%			Mask = 90%			Avg		
		mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1
Fed-DualCoOp	NeurIPS'22	82.52	76.86	76.58	80.52	75.00	74.73	77.65	72.65	69.98	74.43	66.72	63.39	61.28	53.89	50.96	75.28	69.02	67.13
Fed-SCPNet	CVPR'23	78.81	70.83	72.60	73.55	72.04	65.77	70.97	70.06	70.62	67.46	60.09	55.04	64.73	56.51	52.93	71.10	65.91	63.39
Fed-MaPLe	CVPR'23	81.27	75.59	73.49	81.16	75.46	74.11	78.50	76.30	74.11	70.60	65.60	64.75	59.53	58.64	48.62	74.21	70.32	67.02
FedPGP	ICML'24	79.96	73.81	63.81	76.71	66.44	64.05	72.96	67.81	63.81	68.29	66.27	59.70	66.05	60.54	50.65	72.79	66.97	60.40
Fed-TCP	CVPR'24	76.42	76.00	72.50	75.42	74.09	72.54	72.43	70.03	70.59	70.35	68.90	67.86	67.51	62.80	55.90	72.43	70.36	67.88
FedTPG	ICLR'24	84.19	77.45	76.93	82.68	75.31	67.00	82.18	68.19	56.87	80.00	65.85	59.29	68.24	58.68	57.09	79.46	69.10	63.44
Fed-PosCoOp	WACV'25	81.18	76.44	73.29	79.86	76.13	72.96	78.50	64.62	71.73	73.87	65.63	67.74	60.27	58.90	54.65	74.74	68.34	68.07
FedAWA	CVPR'25	80.81	70.70	72.94	75.37	69.16	65.68	75.82	69.66	65.87	72.59	67.45	57.16	68.22	55.27	56.92	74.56	66.45	63.71
Fed-RAM	CVPR'25	87.89	77.03	75.93	85.42	75.78	73.41	82.30	74.84	72.93	79.58	72.67	68.41	68.09	63.18	59.84	80.66	72.70	70.10
FedMVP	ICCV'25	85.27	79.10	76.23	84.34	76.83	75.99	82.09	74.20	73.07	80.64	75.03	70.15	66.53	62.89	58.40	79.77	73.61	70.77
FedMPT	Ours	**89.40**	**83.06**	**81.12**	**87.40**	**82.12**	**81.48**	**88.75**	**80.86**	**79.71**	**86.18**	**81.32**	**74.47**	**74.24**	**66.23**	**64.93**	**85.19**	**78.72**	**76.34**
Δ Prev. Best	\	**+1.51**	**+3.96**	**+4.19**	**+1.98**	**+5.29**	**+5.49**	**+6.45**	**+4.56**	**+5.60**	**+5.54**	**+6.29**	**+4.32**	**+6.00**	**+3.05**	**+5.09**	**+4.54**	**+5.11**	**+5.57**
COCO2014																			
Methods	Venue	Mask = 10%			Mask = 30%			Mask = 50%			Mask = 70%			Mask = 90%			Avg		
		mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1
Fed-DualCoOp	NeurIPS'22	52.44	47.09	56.86	39.56	42.05	37.50	26.55	28.57	22.04	22.72	24.40	21.50	20.96	22.97	12.76	32.45	33.02	30.13
Fed-SCPNet	CVPR'23	50.11	46.30	55.80	38.75	35.24	33.56	23.08	26.73	23.71	24.46	26.82	20.74	21.15	14.82	13.70	31.51	29.98	29.50
Fed-MaPLe	CVPR'23	57.95	54.96	56.32	36.86	39.54	30.13	25.16	27.78	25.22	22.61	24.36	22.82	20.86	18.49	17.23	32.69	33.03	30.34
FedPGP	ICML'24	58.14	51.23	54.25	35.30	38.89	38.32	23.12	29.44	20.36	21.57	28.09	20.84	23.14	16.19	17.15	32.25	32.77	30.18
Fed-TCP	CVPR'24	57.88	50.20	55.15	40.70	41.85	25.07	24.61	27.73	30.54	22.57	24.51	19.57	21.00	18.48	12.57	33.35	32.55	28.58
FedTPG	ICLR'24	60.49	52.85	53.17	42.71	40.08	41.60	21.40	25.88	26.74	23.28	25.32	22.97	20.93	16.55	14.65	33.76	32.14	31.83
Fed-PosCoOp	WACV'25	52.18	49.70	54.06	39.70	40.04	40.58	27.03	23.68	27.23	22.49	24.98	23.67	19.77	13.00	19.57	32.23	30.28	33.02
FedAWA	CVPR'25	58.80	52.55	58.94	38.40	38.06	39.30	28.06	28.80	34.64	21.00	26.81	24.50	18.37	18.16	23.13	32.93	32.88	36.10
Fed-RAM	CVPR'25	60.80	55.85	59.89	42.68	42.82	50.34	29.79	33.25	38.51	24.86	27.63	26.42	20.26	19.52	23.50	35.68	35.81	39.73
FedMVP	ICCV'25	58.80	53.75	55.10	41.64	39.51	45.52	28.46	28.19	36.07	26.26	29.82	25.53	25.82	24.86	24.54	36.20	35.23	37.35
FedMPT	Ours	**62.75**	**55.29**	**61.02**	**45.98**	**44.68**	**54.25**	**33.13**	**36.22**	**44.22**	**30.54**	**33.95**	**32.06**	**30.39**	**32.84**	**30.78**	**40.56**	**40.60**	**44.47**
Δ Prev. Best	\	**+1.95**	**+-0.56**	**+1.13**	**+3.27**	**+1.86**	**+3.91**	**+3.34**	**+2.97**	**+5.71**	**+4.28**	**+4.13**	**+5.64**	**+4.57**	**+7.98**	**+6.24**	**+4.36**	**+4.78**	**+4.73**
NUS-Wide																			
Methods	Venue	Mask = 10%			Mask = 30%			Mask = 50%			Mask = 70%			Mask = 90%			Avg		
		mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1
Fed-DualCoOp	NeurIPS'22	40.13	35.57	61.16	31.20	22.90	43.49	24.79	23.03	28.87	13.04	9.04	16.05	10.81	9.34	4.17	23.99	19.98	30.75
Fed-SCPNet	CVPR'23	35.44	33.22	56.39	28.95	20.45	39.91	23.64	21.99	26.73	13.28	8.96	19.90	9.97	9.61	6.24	22.26	18.85	29.83
Fed-MaPLe	CVPR'23	48.60	40.67	71.28	30.20	25.62	50.50	25.10	22.49	37.35	13.85	11.06	26.01	6.28	5.53	11.20	24.81	21.07	39.27
FedPGP	ICML'24	42.24	36.90	68.55	36.54	32.36	55.74	23.24	26.90	41.55	19.67	12.60	28.03	7.98	4.95	9.54	25.93	22.74	40.68
Fed-TCP	CVPR'24	47.19	40.70	61.81	35.40	35.53	43.67	26.86	26.58	34.40	12.38	12.06	15.34	9.86	9.46	12.66	26.34	24.87	33.58
FedTPG	ICLR'24	47.64	38.60	66.39	34.60	32.17	52.85	26.85	28.76	41.05	20.53	12.23	22.28	11.52	10.47	12.89	28.23	24.45	39.09
Fed-PosCoOp	WACV'25	40.64	39.38	63.42	35.94	34.18	45.25	24.19	24.50	38.49	12.74	10.92	25.42	10.92	10.99	8.03	24.89	23.99	36.12
FedAWA	CVPR'25	46.86	41.91	70.31	36.82	36.65	54.83	25.83	28.62	43.67	17.39	21.82	28.76	12.35	12.18	11.08	27.85	28.24	41.73
Fed-RAM	CVPR'25	45.53	39.83	72.06	37.81	37.01	59.51	26.86	26.07	51.93	20.92	22.03	27.50	11.60	13.91	12.25	28.54	27.77	44.65
FedMVP	ICCV'25	48.06	42.28	71.44	36.06	36.38	59.72	25.61	26.00	45.46	15.04	16.64	28.54	10.18	12.04	10.38	26.99	26.67	43.11
FedMPT	Ours	**51.72**	**45.16**	**72.88**	**40.15**	**40.66**	**63.72**	**30.42**	**32.83**	**56.64**	**24.47**	**26.27**	**31.05**	**15.81**	**18.97**	**15.01**	**32.51**	**32.78**	**47.86**
Δ Prev. Best	\	**+3.12**	**+2.88**	**+0.82**	**+2.34**	**+3.65**	**+4.00**	**+3.56**	**+4.07**	**+4.71**	**+3.55**	**+4.24**	**+2.29**	**+3.46**	**+5.06**	**+2.12**	**+3.97**	**+4.54**	**+3.21**

Table 3. Results on the real-world MLR Benchmark. We report the mAP, CF1, and OF1. The best results are marked with **bold**.

Method	Multi-Scene			MLRSNet		
	mAP	CF1	OF1	mAP	CF1	OF1
Fed-DualCoOp	40.09	31.37	50.18	38.24	40.61	66.97
Fed-MaPLe	33.18	30.02	29.70	37.37	46.37	61.18
FedPGP	45.96	35.85	53.77	52.75	45.27	50.75
Fed-TCP	45.25	37.43	53.43	43.01	42.32	67.61
FedTPG	44.45	35.55	52.39	31.38	35.11	62.65
Fed-PosCoOp	42.30	35.06	47.79	37.25	39.89	65.14
FedAWA	47.58	37.92	53.53	40.89	42.32	68.13
Fed-RAM	49.41	39.07	52.70	47.83	46.07	66.18
FedMVP	49.56	39.62	54.16	45.89	44.98	66.52
FedMPT	**53.68**	**43.97**	**57.83**	**58.76**	**50.86**	**71.22**
Δ Prev. best	**+4.12**	**+4.35**	**+3.67**	**+6.01**	**+4.49**	**+3.09**

under severe heterogeneity, but approaches MLR methods in average metrics for their suboptimal multi-label modeling capabilities. In contrast, our FedMPT indisputably outperforms them by a substantial margin (mAP: 3.84% on VOC2007, 3.01% on COCO2014, 3.36% on NUS-Wide). Meanwhile, FedMPT shows less fluctuation faced with data heterogeneity, highlighting its robustness and superiority.

Results of Federated Part-annotation Benchmark. The results are shown in Table 2. We find that existing SOTAs like Fed-RAM and FedMVP, which excel on fully-annotated data, generally yield much degraded performance when the annotation mask increases, validated by their average decrease of about 20%, 32%, and 38% in three datasets, respectively. We argue that since these methods rely solely on coarse-grained object categories for cross-modal matching, they overemphasize the adaptation of individual prompts to local data distributions, thereby impairing the generalization capability of the global model. In contrast, our FedMPT consistently outperforms existing

Table 4. Ablations on different proposed modules.

CPs	Adapters	OT	Gate	mAP	CF1	OF1	Avg.
✓				87.08	82.90	81.66	83.88
	✓			84.40	78.31	80.04	80.92
✓	✓			87.62	83.05	83.68	84.78
✓		✓		89.35	83.82	82.53	85.23
✓	✓	✓		89.64	83.75	83.72	85.70
✓	✓		✓	87.89	83.34	83.01	84.75
✓	✓	✓	✓	90.10	83.96	84.50	86.19

Table 5. Comparison of computation overhead on VOC2007.

Method	# Total Param.	# Tunable Param.	Training Time	mAP
Fed-PosCoOp	86.60 M	0.02 M	38.93 ms/iter	84.46
Fed-MaPLe	90.14 M	3.56 M	65.25 ms/iter	81.87
FedTPG	90.79 M	4.21 M	59.43 ms/iter	84.23
Fed-RAM	99.60 M	13.02 M	384.51 ms/iter	85.54
FedMVP	87.72 M	1.14 M	75.80 ms/iter	85.61
FedMPT	87.38 M	0.80 M	97.03 ms/iter	90.10

methods with its decomposition of conditions: on average of three datasets, FedMPT surpasses existing best results by about 5.3%, 5.8%, and 4% at three metrics; the benefits from FedMPT show a positive correlation with `mask` (2.26%→7.25% on COCO2014), verifying its robustness.

Results of Federated Real-world MLR Benchmark. In Table 3 we report the results on two real-world MLR datasets. While real-world datasets present more noise and tricky instances, the reliance on multiple conditions makes our FedMPT more robust to potential noise in the data compared to existing methods. Concretely, we observe that FedMPT keeps its superiority and outperforms existing SOTAs by 4.27% mAP on Multi-Scene and 6.01% mAP in MLRSNet, highlighting its promising versatility.

6. Ablation Studies and Discussions

Proposed Modules. The results are shown in Table 4. We find that Condition Prompts (CPs) provide a more substantial performance gain than the adapters alone, underscoring the critical role of explicit condition modeling over mere visual feature adaptation. Moreover, employing OT delivers an average enhancement of 1.44% mAP. However, employing gating without OT has more limited improvement (+0.27% mAP) than that when OT is applied (+2.21% mAP), indicating that the synergistic effects between conditions rely on OT to mediate trade-offs among patches.

Cost Analysis. Table 5 reports the computation overheads. Fed-PosCoOp yields the least learnable parameters but also the worst performance; FedMVP and FedRAM yield comparable performance (85.61% and 85.54% mAP), but at the cost of significantly expanded training time and parameters. Comparably, FedMPT achieves the best performance with minor extra overhead, showing its efficiency.

Number of Learnable Tokens β_{cond} and β_{cls}. β_{cond} and β_{cls} control the number of learnable tokens of conditions and classes, respectively. We perform an exemplar grid-search to them on VOC2007 in Figure 6.a. While we can

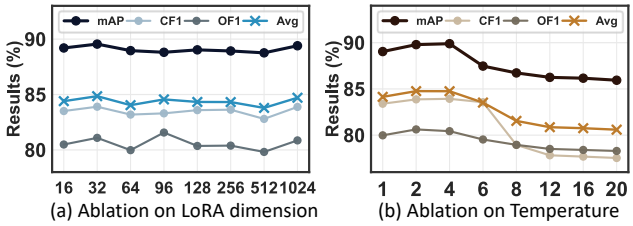

Figure 5. Ablation studies on LoRA dimension and temperature.

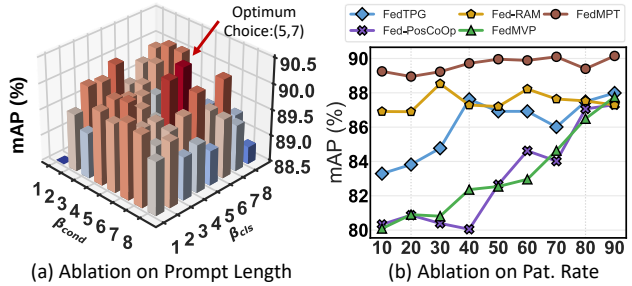

Figure 6. Ablation studies on prompt length and participation rate.

observe that less learnable tokens tend to result in poor performance, excessively large choices also cause minor degradation, where possible reasons lie in the increased learning difficulty of the prompts. The optimal choice is (5,7), which is a trade-off between the capability and difficulty.

LoRA Dimension D_s and Temperature τ. We separately alter D_s and τ to [16-1024] and [1-20] and investigate the mAP in Figure 5. **For D_s**, we find that its change brings a minor fluctuation generally and a continual decrement as D_s grows larger than 32, possibly from overfitting. $D_s = 32$ is the optimal choice. **For τ**, the results show an obvious degradation of all metrics when it grows larger than 4, probability due to the indistinguishability between classes. Our experiments show that $\tau = 4$ is the optimal.

Participation Rate of Clients. We vary the participation rate ϵ from 10% to 90% and report the results in Figure 6.b. We can observe that FedMPT consistently outperforms other methods and exemplifies a gentle change under different participation rates. In contrast, some methods like FedMVP and FedTPG suffer from dramatic degradation (∼5%/7% mAP) when participation rate decreases.

7. Conclusion

The integration of Multi-Label Recognition (MLR) with Federated Learning (FL) introduces a significant risk of overfitting to spurious local label correlations. To address this, we present FedMPT, a novel framework grounded in causal analysis that leverages conditions to approximate true class relationships. FedMPT employs an LLM-driven pipeline to generate abstract conditions, aligns them with visual regions via optimal transport, and integrates their contributions through a gating mechanism. Experiments validate the superiority across federated benchmarks.

References

[1] Rabab Abdelfattah, Qing Guo, Xiaoguang Li, Xiaofeng Wang, and Song Wang. Cdul: Clip-driven unsupervised learning for multi-label image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1348–1357, 2023. 2

[2] Wenxuan Bao, Ruxi Deng, Ruizhong Qiu, Tianxin Wei, Hanghang Tong, and Jingrui He. Latte: Collaborative test-time adaptation of vision-language models in federated learning. *arXiv preprint arXiv:2507.21494*, 2025. 3

[3] Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, 2025. 2

[4] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8991–9000, 2020. 1

[5] Shikai Chen, Jianfeng Wang, Yuedong Chen, Zhongchao Shi, Xin Geng, and Yong Rui. Label distribution learning on auxiliary label space graphs for facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13984–13993, 2020. 1

[6] Shiming Chen, Bowen Duan, Salman Khan, and Fahad Shahbaz Khan. Interpretable zero-shot learning with locally-aligned vision-language model. *arXiv preprint arXiv:2506.23822*, 2025. 4

[7] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 522–531, 2019. 2

[8] Tianlong Chen, Xuxi Chen, Xianzhi Du, Abdullah Rashwan, Fan Yang, Huizhong Chen, Zhangyang Wang, and Yeqing Li. Adamv-moe: Adaptive multi-task vision mixture-of-experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17346–17357, 2023. 2

[9] Zhao-Min Chen, Xiu-Shen Wei, Xin Jin, and Yanwen Guo. Multi-label image recognition with joint class-aware map disentangling and label correlation embedding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 622–627. IEEE, 2019. 1

[10] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5177–5186, 2019. 1, 2

[11] 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. 5

[12] Tianyu Cui, Hongxia Li, Jingya Wang, and Ye Shi. Harmonizing generalization and personalization in federated prompt learning. *arXiv preprint arXiv:2405.09771*, 2024. 1, 2, 3, 5, 6

[13] Wenlong Deng, Christos Thrampoulidis, and Xiaoxiao Li. Unlocking the potential of prompt-tuning in bridging generalized and personalized federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6087–6097, 2024. 3

[14] 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. 5

[15] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 647–657, 2019. 1

[16] 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. 5

[17] Bin-Bin Gao and Hong-Yu Zhou. Learning to discover multi-class attentional regions for multi-label image recognition. *IEEE Transactions on Image Processing*, 30:5920–5932, 2021. 1

[18] Tao Guo, Song Guo, Junxiao Wang, Xueyang Tang, and Wenchao Xu. Promptfl: Let federated participants cooperatively learn prompts instead of models–federated learning in age of foundation model. *IEEE Transactions on Mobile Computing*, 23(5):5179–5194, 2023. 2

[19] Ping Hu, Ximeng Sun, Stan Sclaroff, and Kate Saenko. Dualcoop++: Fast and effective adaptation to multi-label recognition with limited annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3450–3462, 2023. 2

[20] 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. 5

[21] Zilong Huang, Qinghao Ye, Bingyi Kang, Jiashi Feng, and Haoqi Fan. Classification done right for vision-language pre-training. *Advances in Neural Information Processing Systems*, 37:96483–96504, 2024. 1

[22] Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8776–8786, 2020. 2

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[24] Liangze Jiang and Tao Lin. Test-time robust personalization for federated learning. *arXiv preprint arXiv:2205.10920*, 2022. 3

[25] 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. 5

[26] Dongseob Kim and Hyunjung Shim. Classifier-guided clip distillation for unsupervised multi-label classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4661–4671, 2025. 1, 2

[27] Wei-Bin Kou, Qingfeng Lin, Ming Tang, Sheng Xu, Rongguang Ye, Yang Leng, Shuai Wang, Guofa Li, Zhenyu Chen, Guangxu Zhu, et al. pfedlvm: A large vision model (lvm)-driven and latent feature-based personalized federated learning framework in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2025. 3

[28] Hongxia Li, Wei Huang, Jingya Wang, and Ye Shi. Global and local prompts cooperation via optimal transport for federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12151–12161, 2024. 2, 3

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[31] Zheng Li, Yibing Song, Ming-Ming Cheng, Xiang Li, and Jian Yang. Advancing textual prompt learning with anchored attributes. *arXiv preprint arXiv:2412.09442*, 1, 2024. 4

[32] 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. 5

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[34] Yingqi Liu, Qinglun Li, Jie Tan, Yifan Shi, Li Shen, and Xiaochun Cao. Understanding the stability-based generalization of personalized federated learning. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[35] Wang Lu, Xixu Hu, Jindong Wang, and Xing Xie. Fedclip: Fast generalization and personalization for clip in federated learning. *arXiv preprint arXiv:2302.13485*, 2023. 3

[36] Leilei Ma, Hongxing Xie, Lei Wang, Yanping Fu, Dengdi Sun, and Haifeng Zhao. Text-region matching for multi-label image recognition with missing labels. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6133–6142, 2024. 2

[37] Lei-Lei Ma, Shuo Xu, Ming-Kun Xie, Lei Wang, Dengdi Sun, and Haifeng Zhao. Correlative and discriminative label grouping for multi-label visual prompt tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25434–25443, 2025. 1, 2, 5

[38] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 1, 2, 5

[39] Kevin Miller, Aditya Gangrade, Samarth Mishra, Kate Saenko, and Venkatesh Saligrama. Sparc: Score prompting and adaptive fusion for zero-shot multi-label recognition in vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4313–4321, 2025. 1, 2

[40] Sanath Narayan, Akshita Gupta, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Mubarak Shah. Discriminative region-based multi-label zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8731–8740, 2021. 2

[41] 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. 5

[42] 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, 6

[43] 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. PmLR, 2021. 1

[44] 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. 1, 2, 3, 5

[45] Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan L Yuille. Multiple instance visual-semantic embedding. In *BMVC*, 2017. 2

[46] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 82–91, 2021. 1, 3

[47] 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, 6

[48] 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. 1, 2, 3, 6

[49] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967. 4

[50] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017. 1, 2

[51] 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, 5

[52] 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. 2, 5

[53] Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Cheng-Zhong Xu. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *Advances in Neural Information Processing Systems*, 37:9565–9584, 2024. 4, 5

[54] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2016. 2

[55] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. Multi-label classification with label graph superimposing. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12265–12272, 2020. 1

[56] Jie Wen, Yicheng Liu, Chao Huang, Chengliang Liu, Yong Xu, and Xiaochun Cao. Causal interventional prompt tuning for few-shot out-of-distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

[57] Qiong Wu, Zhaoxi Ke, Yiyi Zhou, Xiaoshuai Sun, and Rongrong Ji. Routing experts: Learning to route dynamic experts in existing multi-modal large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 5

[58] Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23826–23837, 2024. 4

[59] 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. 5

[60] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. Cross-modality attention with semantic graph embedding for multi-label classification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12709–12716, 2020. 1

[61] Hao Yu, Xin Yang, Xin Gao, Yan Kang, Hao Wang, Junbo Zhang, and Tianrui Li. Personalized federated continual learning via multi-granularity prompt. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4023–4034, 2024. 3

[62] Yanan Zhang, Jiangmeng Li, Lixiang Liu, and Wenwen Qiang. Rethinking misalignment in vision-language model adaptation from a causal perspective. *Advances in Neural Information Processing Systems*, 37:39224–39248, 2024. 3

[63] Yuqi Zhang, Xiucheng Li, Hao Xie, Weijun Zhuang, Shihui Guo, and Zhijun Li. Multi-label action anticipation for real-world videos with scene understanding. *IEEE Transactions on Image Processing*, 33:3242–3255, 2024. 1

[64] Yifei Zhang, Hao Zhu, Alysa Ziying Tan, Dianzhi Yu, Longtao Huang, and Han Yu. pfedmxf: Personalized federated class-incremental learning with mixture of frequency aggregation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30640–30650, 2025. 3

[65] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018. 1, 2

[66] 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. 6