electronics

MDPI

Article

A Federated CLIP Fine-Tuning Method Based on Optimal Transport and Dual Prompt Personalization

Lei Shi, Zepeng Li *, Xu Ding, Yingfei Zhu and Xin Yao

School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China; shilei@hfut.edu.cn (L.S.)

* Correspondence: 2023110578@mail.hfut.edu.cn

Abstract

The Contrastive Language-Image Pre-training (CLIP) model uses contrastive learning to align image and text representations, and fine-tuning CLIP with federated learning can extend its application to professional fields. However, federated CLIP fine-tuning faces two key challenges: insufficient alignment of fine-grained semantics between vision and text modalities and poor adaptability to non-independent and identically distributed (non-IID) data. This paper proposes the Optimal Transport Dual Prompt Personalization (OTDPP) framework, injects prompt parameters into the deep networks of both visual and text encoders, achieves fine-grained cross-modal alignment through optimal transport, and designs a dual prompt tuning mechanism. The framework splits prompt parameters into a shared global part aggregated by the server and a private local part reserved by clients, and it enables personalized adaptation without updating large backbone encoders. Extensive experiments show that compared with classic prompt tuning baseline methods, OTDPP reduces computational and communication overhead, retains client-specific personalized features, significantly improves model adaptability and performance, and thus demonstrates broad application prospects.

Keywords: federated learning; prompt tuning; CLIP

1. Introduction

Machine learning [1,2] continuously improves generalization and predictive capabilities by automatically learning patterns from data. Its evolution from supervised learning to self-supervised learning lays the foundation for more versatile intelligent models. With the rapid development of artificial intelligence technology, large models based on the Transformer [3] architecture have achieved significant breakthroughs in the interdisciplinary field of vision and language. Among them, the Contrastive Language-Image Pre-training (CLIP) [4] model has significantly improved cross-modal matching and semantic understanding capabilities between images and text by conducting contrastive learning on massive image–text pairs. Unlike traditional supervised learning methods relying on fixed category labels, CLIP uses natural language as supervision signals, enables the model to directly capture deep semantic correlations between images and text, and thus exhibits strong generalization performance in tasks such as zero-shot image recognition and cross-modal retrieval. This mechanism provides a new paradigm for the unified modeling of vision–language tasks. Relevant studies have confirmed that CLIP has wide applications in various visual tasks including image classification, object detection and

Academic Editors: Zhipeng Cai, Sai Akshita Maradapu Vera Venkata and Chenyu Wang

Received: 5 January 2026
Revised: 11 February 2026
Accepted: 25 February 2026
Published: 27 February 2026

even multimodal generation, and strongly promotes the evolution and development of general visual systems.

When applying such large models to professional fields, there are currently two main deployment paths. First is calling cloud-based large model services—accessing large models hosted by third parties through APIs or web interfaces. This method is convenient but may pose risks of local user data leakage, especially in fields with strict privacy requirements such as medical care and finance, and this problem becomes more prominent. For example, the European Union has issued strict regulations such as GDPR [5] to address data security challenges and strengthen personal information protection. In addition, cloud-based general models often lack precision in vertical scenarios, struggle to provide personalized services based on users' local data, and lack deep logical reasoning capabilities for professional tasks. Second is fine-tuning [6] large models locally with high-quality domain data to make them more in line with professional needs. But this approach is highly dependent on data scale and quality. With limited samples, the model may fail to fully learn knowledge in subdivided domains, leading to restricted application effects.

Therefore, in the current process of deploying large models in professional fields [7], we urgently need to systematically address a series of challenges including data privacy protection, insufficient personalized services, and difficulty in knowledge transfer under few-shot scenarios. Federated learning [8] is a classic distributed machine learning paradigm. Multiple clients collaboratively train the same global model and share only model parameters instead of local private data—this approach greatly protects local data privacy and alleviates the problem of insufficient data samples as different clients train their own local data. Thus, scholars regard the combination of large model fine-tuning and federated learning as a promising solution to these current problems in the field. Currently, in the field of model fine-tuning, an increasing number of models have begun to use federated learning for optimization.

The core goal of traditional federated learning is to train a unified global model through distributed collaboration and achieve cross-client knowledge sharing and performance improvement [9–12]. However, this framework often overlooks the widespread data heterogeneity among participating clients. This heterogeneity specifically manifests in significant differences in local data of different clients in terms of data distribution, class proportion, task preference and other aspects. For instance, image data from different users or institutions may vary significantly in shooting scenarios and lighting conditions, while text data may differ greatly in description styles and domain terminology usage. These differences lead to a significant drop in generalization performance of the global model trained by traditional federated learning on local tasks of some clients, making it hard to meet practical application requirements.

To address the dilemma caused by this data heterogeneity, introducing personalized federated learning [13] has become a natural and inevitable development direction. Recent advances in personalized federated learning explore various technical paths, including model-based personalization, model mixing, and personalized model architectures. Unlike traditional federated learning, the core demand of personalized federated learning is, on the premise of ensuring distributed collaboration and data privacy, not only to learn a global model with strong common knowledge but also to tailor a personalized model for each client that adapts to its local data characteristics. It achieves an organic balance between global collaboration advantages and local data adaptability, thus breaking through the performance bottleneck of traditional federated learning in heterogeneous data scenarios.

Although the concept of personalized federated learning has been widely recognized, most current federated large model fine-tuning studies still have obvious shortcomings. They fail to fully consider and meet the personalized needs of clients, and thus the adapta-

tion effect of related methods in heterogeneous data scenarios is limited. To address this issue, this paper proposes a personalized prompt tuning method based on dual prompts and optimal transport strategy. The introduction of the optimal transport strategy aims to efficiently align feature representations of visual and text modalities, optimize the matching degree of cross-modal feature distributions, and help the model achieve a better balance between task adaptation capability and generalization capability. Meanwhile, the dual prompt mechanism can specifically capture common knowledge across clients and personalized features of individual clients, effectively address the data heterogeneity problem among clients, and improve the adaptation performance of personalized models.

The main contributions of this paper can be summarized as follows:

- This paper proposes a method combining personalized federated learning and Prompt tuning to solve the challenges of data privacy protection, insufficient personalized services, and difficulty in few-shot knowledge transfer during the deployment of large models in professional fields.
- We introduce the optimal transport strategy to optimize cross-modal feature learning of the CLIP model.
- Comprehensive experiments are conducted on simulated heterogeneous datasets for federated cross-modal tasks, verifying the advantages of the proposed framework over existing methods.

2. Related Work

2.1. Contrastive Language-Image Pre-Training Model

The Transformer architecture serves as the fundamental framework for modern large-scale models. It innovatively adopts the self-attention mechanism to capture connections between all elements in the input sequence and uses positional encoding to preserve order information. With exceptional flexibility and scalability, the Transformer has driven breakthroughs across various fields of artificial intelligence. In computer vision, Vision Transformers (ViTs) have demonstrated that global attention can effectively model image structures. In multimodal AI, it has further become a unified architecture to align heterogeneous data such as text, images and audio and perform joint reasoning. The Transformer has evolved into a universal foundational architecture for contemporary AI systems.

Building on this foundation, Vision–Language Models (VLMs) such as CLIP have achieved unified vision and text understanding by jointly training independent image and text encoders. These encoders map the two modalities into a shared embedding space and align corresponding image–text pairs within this space. This cross-modal alignment supports multiple application scenarios, including image caption generation, visual question answering, and context-aware systems such as autonomous driving.

CLIP initially trained on 400 million image–text pairs. Key findings show that optimizing text descriptions significantly improves zero-shot classification accuracy, and this has driven the development of subsequent tuning methods such as CoOp. CLIP's zero-shot capability enables the identification of unseen classes by comparing image features with text category descriptors. While subsequent studies have improved prompt learning in few-shot and unsupervised scenarios, these methods usually require full-model backpropagation and result in high computational costs. Notably, the field of adapting CLIP in federated learning frameworks to leverage its alignment potential under decentralized data remains under intensive exploration.

2.2. Personalized Federated Learning

The core goal of personalized federated learning is to learn personalized models adapted to the characteristics of local data for different clients within the federated frame-

work. Based on different technical paths, existing personalized federated learning studies can be mainly divided into the following categories. Model-based personalization methods perform regularization or fine-tuning of the global model locally on clients to adapt to local data distributions. Model mixing methods do not rely solely on a single global model but customize model combinations for each client. For example, clustered federated learning [14] groups similar clients into the same cluster, each cluster has a distinct global model, and each client achieves personalization through weighted combination. Personalized model architecture methods introduce personalized components at the model design level. For instance, DP2FL [15] splits the model into globally shared layers and local personalized layers. Shared layers learn general features across clients, while personalized layers focus on capturing locally specific patterns. DP2FL is not designed for multimodal learning and lacks explicit cross-modal alignment mechanisms. In contrast, OTDPP injects prompts into deep layers of both encoders and retains personalized prompts locally to adapt to non-IID data. This enables superior personalization and cross-modal understanding.

In recent years, federated prompt learning has made significant progress in the personalized fine-tuning of vision–language models. pFedMMA [16] proposes a personalized federated learning framework based on multimodal adapters. It decomposes adapters into modality-specific up-down projection layers and globally shared projection layers, and only aggregates shared components. This design achieves a better balance between generalization and personalization under label and feature shifts while maintaining communication efficiency. pFedMoAP [17] introduces a federated prompt learning framework based on Mixture of Experts (MoE). It allows clients to download pre-aggregated prompts as non-local experts and dynamically fuse expert knowledge using local attention gating networks. This method achieves efficient cross-client knowledge sharing and improved personalized performance under extreme data heterogeneity.

To better evaluate the effectiveness of personalized federated learning frameworks, researchers simulate data distribution heterogeneity in the real world and establish benchmark datasets with heterogeneous data partitions in the field of image classification. Classic methods adopt Dirichlet distribution to partition original data, and the datasets use benchmark datasets that can represent various scenarios. As shown in Figure 1, we take the Oxford Flowers-102 dataset as an example to demonstrate the highly non-IID partitions generated by Dirichlet distribution (α = 0.1), where some clients may completely lack samples of certain categories. This partitioning strategy is extended to all datasets in our experiments to ensure consistent simulation of non-IID scenarios.

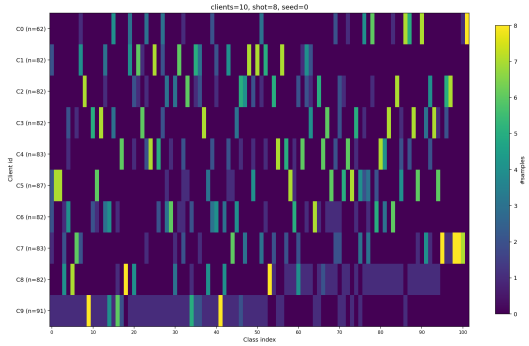

Figure 1. Visualization of Dirichlet Distribution (α = 0.1) on Non-IID Oxford Flowers-102 Dataset. This partitioning strategy is uniformly applied to all benchmark datasets in the experiments.

2.3. Optimal Transport Theory

Optimal transport was originally developed to find a transport plan for moving mass from the source distribution to the target distribution. Given the source distribution, target

distribution, and a transport cost function defined on the sample space, the method minimizes the total expected cost. Optimal transport theory provides a powerful and rigorous mathematical framework. It measures differences between probability distributions and plans the minimum cost of mass transport between them. In recent years, this theory has demonstrated great value in multiple directions such as domain adaptation, generative models, and cross-domain learning [18].

Recently, Li et al. proposed FedOTP [19]. This method adopts a dual-prompt structure and unbalanced optimal transport to achieve federated fine-tuning of the CLIP model. FedOTP balances global consensus and local personalization by aligning visual features with prompts. However, OTDPP differs architecturally by introducing a hierarchical injection deep prompt learning strategy into visual and text encoders. This achieves fine-grained alignment across multiple network layers. However, the key characteristic of FedOTP lies in its use of unbalanced Optimal Transport to synergistically align global and local prompts, enabling prompts to focus solely on critical image patches rather than the entire content, thereby effectively balancing consensus and personalization. In addition, OTDPP's personalization mechanism retains local prompts entirely on clients without server aggregation. It uses unbalanced optimal transport to achieve client-specific adaptation for non-independent and identically distributed data, while FedOTP focuses on global–local prompt collaboration via optimal transport. Therefore, OTDPP achieves superior personalization performance by combining deep prompt tuning with client-exclusive retention mechanisms.

Its application in Prompt tuning is an important extension of the frontier of optimal transport [20,21]. OT can regard features from text encoders and visual encoders as different distributions. It then uses OT loss to align their statistical characteristics and thus improve model performance.

2.4. Multi-Modal Prompt Fine-Tuning

Prompt tuning, as a parameter-efficient fine-tuning paradigm, has expanded from the traditional field of natural language processing to the field of vision–language models. Its core idea is to freeze the backbone network of pre-trained VLMs, add only a small number of learnable prompt parameters, and guide the model to adapt to new tasks. The core of VLM prompt learning lies in cross-modal prompt design, and existing prompt designs can be divided into text-driven, vision-driven, and cross-modal hybrid prompts [22].

Text prompts are the most commonly used form in early prompt learning, building task-related templates based on natural language. Khattak et al. [23] propose training prompts using only text data generated by large language models, distill contextual knowledge through language-specific efficient training schemes, achieve zero-shot transfer of prompts to new categories and datasets, and reduce prompt engineering costs. The advantages of text prompts lie in strong flexibility and compatibility with most VLMs. Vision prompts inject task-related signals into the visual encoders of VLMs. For example, Franklin et al. [24] propose the VDPO framework, integrate visual embedding prompt tuners, text instruction generators, and visual generation modules, dynamically generate text prompts from visual inputs, guide high-fidelity image synthesis, and improve generation performance in multiple benchmark tests. Vision prompts enhance task-related visual perception capabilities while relying heavily on the structure of visual encoders and have poor cross-model generality.

Hybrid prompts that combine text and vision prompts are the current research trend. Zheng et al. [25] propose the Hicropl framework, establish bidirectional knowledge flow between text and visual modalities through hierarchical knowledge mappers, mutually optimize semantics at different network depths, integrate consistency regularization, and

enhance cross-modal alignment effects and generalization capabilities of vision–language models. Hybrid prompts solve the problems of unimodal prompts when the design of bimodal fusion strategies becomes more complex. In addition, deep prompt injection is another research hotspot in current prompt tuning. It inserts prompts into multiple layers of the Transformer, allows task adaptation signals to penetrate the deep layers of the model, and thus adjusts the multi-level representations of the model more finely.

In the context of federated learning, multi-modal prompt fine-tuning has emerged as a promising direction to address data privacy and domain adaptation challenges, but existing works still have notable limitations. For instance, PromptFL [26] proposes a federated prompt learning paradigm where participants collaboratively learn prompts instead of full models, but it focuses solely on single-modal adaptation and neglects cross-modal semantic alignment. TPG [27] introduces dual text and visual prompts for federated scenarios but lacks personalized design for non-IID data and fails to achieve fine-grained cross-modal matching. DP2FL [15], a personalized federated learning method for foundation models, utilizes dual prompts for global and local adaptation, inherently involves cross-modal feature alignment through its base model CLIP, and personalizes models via prompt-based aggregation instead of gradient decoupling. These works highlight the urgent need for a unified framework that integrates fine-grained cross-modal alignment and personalized adaptation in federated multi-modal prompt learning.

However, current research still faces some challenges: designing a unified framework to simultaneously optimize multimodal prompts and ensure their coordination, and improving model generalization capabilities under few-shot scenarios.

3. Approach

3.1. Problem Setup

In federated learning scenarios based on large-scale vision–language pre-trained models such as CLIP, N clients hold non-iid private data $\mathcal{D}_i = \{(x_j, y_j)\}_{j=1}^{n_i}$. Here, n_i is the number of data samples of client i, and all clients aim to collaboratively train a model adapted to specific downstream tasks. As a core paradigm of parameter-efficient fine-tuning, multimodal prompt learning has found wide application in vision–language models such as CLIP.

However, applying cross-modal prompt learning in federated scenarios faces unique and yet unresolved challenges. While data distributions across different clients show significant differences, traditional global feature alignment methods struggle to capture fine-grained vision–text correlations, and this leads to degraded model performance on personalized tasks. Existing federated multi-modal prompt methods such as TPG [27] and PromptFL [26] either ignore personalized division or lack cross-modal fine-grained alignment.

We thus propose the Optimal Transport Dual Prompt Personalization (OTDPP) algorithm. This method injects learnable prompt tokens into the multi-layer networks of visual and text encoders and achieves collaborative optimization of local features and text descriptions. It also introduces optimal transport theory to enhance fine-grained alignment of vision–text features and designs a hierarchical prompt tuning strategy, decomposes prompt parameters into a globally shared part and a locally private part. Without fine-tuning the backbones of visual and text encoders, we only update a small number of prompt-related parameters, reduce computational and communication overhead, and allow clients to retain personalized capabilities. The core goal of the OTDPP algorithm is to learn a set of globally shared prompt parameters in federated learning and reserve a portion of private prompt parameters for each client to adapt to its local distribution. The main system framework is shown in Figure 2.

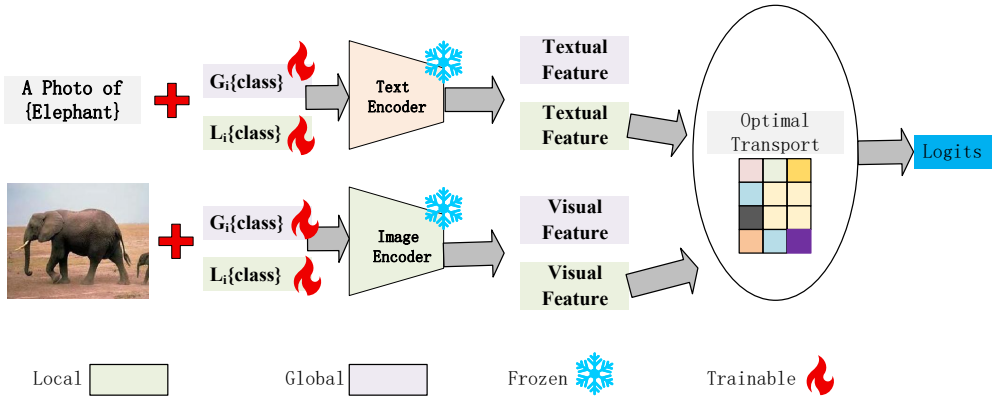

Figure 2. The personalized federated learning process of the OTDPP framework, demonstrating global prompt aggregation and local prompt retention between the server and clients.

In the following subsections, we elaborate on the design principles and implementation methods of these components in detail.

3.2. Prompt Tuning Method for the CLIP Model

In this subsection, we first introduce the working principle of CLIP. CLIP consists of two core components: the image encoder $I(\cdot)$ and the text encoder $T(\cdot)$. The image encoder outputs token features of images, and the text encoder outputs text features of categories. For an input image x and a category text t_c, CLIP first calculates the normalized image feature $f_i = I(x)$ and text feature $f_t^{(c)} = T(t_c)$, then obtains the similarity score $s_c = f_i \cdot f_t^{(c)}$ through cosine similarity. The features are normalized and the dot product equals cosine similarity.

These scores are converted into predicted probabilities for each category through the softmax function, with the formula as follows:

$$p(y = c|x) = \frac{\exp(s_c/\tau)}{\sum_{j=1}^{C} \exp(s_j/\tau)}, \tag{1}$$

here, τ is the temperature parameter, and C is the total number of categories.

In prompt tuning, we optimize text feature representations by designing prompt texts. We use the template "a photo of a class" to generate category-specific prompt texts p_c. Similarly, the text encoder encodes prompt texts to obtain features $f_p^{(c)} = T(p_c)$, and calculates the similarity score $s'_c = f_i \cdot f_p^{(c)}$ with image features. Finally, we use the above formula to calculate predicted probabilities. Through this approach, prompt tuning can improve the model's classification performance on specific tasks.

3.3. Optimal Transport

In the proposed OTDPP framework, balanced optimal transport is uniformly applied. It strictly obeys the mass conservation constraint, ensuring complete mass matching between visual and text feature sets and avoiding instability caused by partial transport. Unlike the FedOTP algorithm, balanced OT can be efficiently optimized via the Sinkhorn algorithm, which facilitates fine-grained alignment between multi-prompts and local visual features. This prevents prompts from collapsing to the same representation and enables more comprehensive and robust category characterization.

The goal of optimal transport is to align image tokens with two types of prompts via optimal transport, achieved by learning a set of learnable prompt vectors. Given an input

image x, we first extract the global visual feature v and a set of local visual features $\{v_i\}_{i=1}^{M}$ using CLIP's visual encoder. Here, $M = H \times W$, where H and W denote the height and width of the feature map, respectively. For the k-th class, we initialize N learnable local prompt vectors $\{p_{k,j}\}_{j=1}^{N}$. We define distributions a and b as uniform weights over visual and prompt features, respectively.

After processing with the visual and text encoders, we obtain the set of local visual features $V = [v_1, \ldots, v_M]^\top \in \mathbb{R}^{M \times F}$ and the set of prompt features $P_k = [p_{k,1}, \ldots, p_{k,N}]^\top \in \mathbb{R}^{N \times F}$, where F is the feature dimension. The cost matrix $O = 1 - VP_k^\top$ represents the cosine distance between V and P_k. We choose cosine distance as the cost metric because CLIP's features are normalized, and cosine distance can effectively measure the similarity between normalized features.

Subsequently, optimal transport finds the optimal transport plan T and the corresponding OT distance $D_{OT}(k)$ by optimizing the following objective function:

$$D_{OT}(k) = D_{OT}(a, b \mid 1 - VP_k^\top). \tag{2}$$

The D_{OT} can be interpreted as the minimum cost required to align the local image feature distribution with the prompt feature distribution. The cost matrix O measures the similarity between visual tokens and text tokens using cosine distance. The optimal transport plan T then finds a coupling that minimizes the total cost of transporting mass from one distribution to the other. The resulting $D_{OT}(k)$ from Equation (2) quantifies the alignment quality for class k.

We further use the D_{OT} to compute class prediction probabilities via a softmax function:

$$q_{ot}(y = k \mid x) = \frac{\exp\big((1 - D_{OT}(k))/\tau\big)}{\sum_{k'=1}^{C} \exp\big((1 - D_{OT}(k'))/\tau\big)}. \tag{3}$$

After determining the transport plan T, we update the prompt vector parameters using cross-entropy loss:

$$\mathcal{L} = -\frac{1}{\mathcal{D}} \sum_{x \in \mathcal{D}} \sum_{k=1}^{C} 1_{y=k} \log q_{ot}(y = k \mid x), \tag{4}$$

where y is the true class label of the sample, $\mathcal{D}$ is the training dataset, x is the sample and y is the label. $1_{y=k}$ is the indicator function, which equals 1 when the sample label y is equal to class k, and 0 otherwise. Finally, the classification logits are obtained.

3.4. Multi-Modal Deep Prompt Learning

To efficiently fine-tune the CLIP model for professional domains, our framework adopts a vision–text dual-modal deep prompt tuning method. Prompts act on the text branch and are injected into the vision branch in a specific manner, simultaneously influencing the understanding of category text and the parsing of image content. We introduce learnable prompts across multiple network layers to enhance contextual understanding.

Prompt information is converted into additional prompt tokens, which the vision encoder can utilize. Our framework concatenates these tokens with the visual feature sequence extracted from the input image. This design enables the model to perceive local region features of the image itself while focusing on a set of task-related prompt cues, allowing visual evidence to prioritize discriminative content for target categories.

We introduce q learnable prompt tokens $\dot{T}_k \in \mathbb{R}^{d_v}{}_{k=1}^{q}$ into the visual branch of CLIP, where d_v is the feature dimension of the visual encoder. Each prompt token $\dot{T}_k$ consists of two disjoint sub-components: a global prompt sub-token G_k that captures cross-client shared knowledge, and a local prompt sub-token L_k that encodes client-specific personalized features, i.e., $\dot{T}_k = [G_k; L_k]$. The global sub-components are aggregated across clients,

while the local sub-components are retained by individual clients to maintain personalized adaptation. These integrated prompt tokens are concatenated with the image input tokens and processed jointly.

Within the image encoder L, which is a deep Transformer, the prompt tokens are integrated layer-by-layer.

To clearly distinguish shallow feature extraction and deep semantic alignment stages, we define the first J layers ($1 \leq n \leq J$) as the shallow injection stage and mainly fuse local visual features in this stage. The remaining layers ($J + 1 \leq n \leq U$) form the deep injection stage and enhance cross-modal alignment through continuous propagation of prompt tokens. We unify all specific formulas as follows:

$$[h_n, F_n, \dot{T}_n] = L_n([h_{n-1}, F_{n-1}, \dot{T}_{n-1}]), \quad n = 1, 2, \ldots, U, \tag{5}$$

where L_n is the m-th layer Transformer encoder function, J is the prompt injection layer, h_n denotes the class token, F_n denotes the set of image patch tokens, and $\dot{T}_{n-1}$ represents the prompt tokens from the previous layer.

The division between shallow and deep layers depends on the difference of feature abstraction levels in Transformer networks. Shallow layers mainly capture local features, while deep layers focus more on global semantic information. We inject prompts into the first J layers and guide the model to focus on task-related local patterns in shallow layers. We keep prompt propagation in deep layers and strengthen the correlation between category semantics and visual features.

Finally, we perform linear projection on the class tokens output by the encoder's final layer to obtain highly task-adapted image feature representations. In the text branch, deep text prompts adopt a homologous deep hierarchical injection strategy, similarly integrating prompt tokens gradually across all layers of the Transformer text encoder. This achieves collaborative optimization of text semantic encoding and visual feature encoding.

Through this hierarchical integration mechanism, prompt tokens can gradually guide visual attention to focus on task-related patterns. This enables the visual encoder to achieve effective adaptation even under heterogeneous data distributions. The dual design of cross-modal prompt injection and deep prompt injection proposed in this study establishes an efficient bimodal collaborative adaptation framework, providing core technical support for the accurate fine-tuning of the CLIP model in federated learning scenarios for professional fields.

3.5. Local Training

Our framework employs a dual-prompt mechanism, where the learnable prompt vector is split into two components with a 1:1 ratio between global and personalized prompts. Global Prompts encapsulate task commonalities across clients and enable global knowledge sharing through server-side aggregation. Personalized prompts capture the local data characteristics of each client and facilitate personalized adaptation through local aggregation.

This 1:1 ratio equips the model with both global generalization capability and local adaptation capability. It prevents performance degradation on local data (which can result from relying solely on the global model) and avoids the loss of collaborative efficiency caused by complete localization.

The integrated prompt tokens $\dot{T}_k$ in the deep prompt injection process are decomposed into global and local sub-components. For client i at communication round t, the complete prompt vector is formulated as follows, where G_i^t and L_i^t correspond to the global and local sub-tokens of $\dot{T}_k$ for client i:

$$P_i^t = [G_i^t; L_i^t], \tag{6}$$

here, G_i^t represents the global head, and L_i^t denotes the personalized tail. The placeholder tail is randomly initialized but remains unchanged during aggregation. It only serves to maintain the model structure and is overwritten by client-specific tails in local training, ensuring consistency without performance loss.

At the beginning of each training round, client i downloads the latest global model parameters from the server. Only the global head G_i^t is initialized with the server-provided parameters, while the personalized tail L_i^t is retained from the previous local round (i.e., $L_i^t = L_i^{t-1}$).

During local training, the client updates the entire prompt vector P_i^t by minimizing the OT-based loss function $\mathcal{L}_{\text{OT}}$ on its local dataset $\mathcal{D}_i$. The gradient descent step is expressed as:

$$P_i^{t+1} \leftarrow P_i^t - \eta \nabla \mathcal{L}_{\text{OT}}(P_i^t; \mathcal{D}_i), \tag{7}$$

where η is the learning rate and $\mathcal{L}_{\text{OT}}$ is the loss function based on optimal transport. After local training, the client detaches the updated personalized tail L_i^{t+1} and stores it locally. Only the global head G_i^{t+1} is uploaded to the server for aggregation.

3.6. Global Aggregation

The server aggregates the global heads uploaded by the clients. Let N be the total number of clients participating in round t and n_i denote the number of local data samples of client i. The server computes the weighted average of the global heads as follows:

$$G^{t+1} = \sum_{i=1}^{N} \frac{n_i}{\sum_{j=1}^{N} n_j} G_i^{t+1}. \tag{8}$$

The personalized tail L_i^{t+1} is not aggregated on the server to preserve client-specific knowledge. To ensure structural consistency of global prompt vectors during parameter transmission, the server appends a placeholder tail to the aggregated global head G^{t+1}. The placeholder tail is a dummy vector with exactly the same dimension as the personalized tail L_i^t. It is introduced solely to form a complete prompt structure for the server-to-client parameter distribution, avoiding structural mismatches that would cause errors during data parsing on the client side. The placeholder tail is a fixed, non-trainable vector and carries no task-related semantics. It never participates in gradient descent, parameter update, or feature learning throughout the federated training process. After receiving the global prompt structure from the server, the client immediately discards the placeholder tail and replaces it with the locally preserved personalized tail L_i^t. This operation reconstructs the complete local prompt vector $P_i^t = [G^{t+1}; L_i^t]$ for the current local training round, and ensures client-specific knowledge is never uploaded or aggregated by the server.

This mechanism ensures model architecture consistency during parameter transmission and does not centralize or dilute the personalized features locally preserved by each client.

4. Experiments

4.1. Datasets

In the empirical study on personalized federated learning and prompt tuning for the CLIP model, to evaluate the method's generalization ability and personalized performance under few-shot and non-iid client data distributions, we select five widely used benchmark datasets for image classification: Caltech-101, DTD (Describable Textures

Database), EuroSAT (satellite images), FGVC-Aircraft (fine-grained aircraft classification), and Oxford Flowers-102.

These datasets cover multiple types of visual tasks including general objects, texture categories, remote sensing images, fine-grained classification, and flower recognition, have high category diversity and significant differences in image content and can comprehensively verify the model's adaptability and robustness under different distributions and difficulty levels.

For each dataset, we first randomly split it into training, validation, and test sets at a ratio of 70%–15%–15% and keep the sample proportion of each category in each split consistent with the original dataset to ensure the fairness of evaluation. To simulate the scarcity of client data in real federated scenarios, we further downsample the training set: randomly select eight labeled images from each category to form a global few-shot training set with a total size of $8 \times C$, where C is the number of categories in the dataset. This design significantly limits the total number of global training samples to focus on the model's fine-tuning and personalized capabilities under few-shot learning.

4.2. Implementation Details

To simulate the data distribution differences among clients in real scenarios, we adopt category-level Dirichlet distribution and allocate the above global few-shot training set to K federated clients. Specifically, for each category, we sample a K-dimensional proportion vector from Dirichlet ($\alpha = 0.1$), and then allocate the eight images of this category to each client according to the proportion. Since $\alpha = 0.1$ makes the sampling results highly non-uniform, the data distribution of each client at the category level shows significant differences and forms realistic non-iid data partitioning.

To comprehensively evaluate the proposed method, this study selects six classic algorithms in the field of CLIP prompt tuning as comparative baselines. FedPGP [28] is a federated prompt learning framework guided by CLIP knowledge. This framework jointly optimizes global prompts and personalized prompts and improves model generalization in data-heterogeneous scenarios. This method introduces contrastive loss and low-rank adaptation mechanisms at the prompt level and effectively addresses the generalization–personalization trade-off of traditional federated learning on non-independent and identically distributed data. ProGrad [29] prevents catastrophic forgetting of pre-trained knowledge by projecting fine-tuning gradients. Maple [22] introduces a multi-modal prompt learning approach that jointly learns deep prompts in both vision and language branches of CLIP, coupled via a vision–language prompt coupling function to ensure mutual synergy for more complete adaptation to downstream tasks. TPG [27] introduces dual text and visual prompt mechanisms and optimizes the context of the two modalities separately but did not take into account the personalized needs of the model. PromptSRC [30] enhances out-of-distribution robustness through self-consistency regularization. These methods are widely recognized as standard baselines in this field. FedOTP [19] introduces a federated learning framework that jointly learns global prompts and local prompts, leveraging unbalanced Optimal Transport to align visual and textual features for systematically addressing data heterogeneity.

In the experiments, we adopt a fixed structure of two prompts, one for global aggregation and the other for client personalization. The global communication rounds of federated learning are set to 100, local training rounds to 1, local learning rate to 0.005, and batch size to 8, and we use the SGD optimizer. The backbone network employs ViT-B/16. All experimental results are based on three independent runs with different random seeds.

4.3. Performance Comparison

The results in Table 1 show that the proposed algorithm consistently achieves the best performance across all five datasets. OTDPP maintains a stable advantage on relatively saturated natural image datasets (Caltech101 and Flowers), and it performs particularly prominently on benchmarks with larger domain shifts and higher fine granularity. The average performance across the five datasets reaches 75.73%. This indicates strong cross-dataset generalization ability under the same experimental conditions.

Table 1. Performance Comparison under Dirichlet Distribution (α = 0.1).

	Caltech101	**DTD**	**Eurosat**	**Aircraft**	**Flowers**
OTDPP	95.00	79.26	78.75	38.88	86.76
FedPGP	92.75	76.86	61.25	31.83	84.82
ProGrad	94.73	56.68	46.40	30.66	80.11
Maple	94.97	59.04	47.05	32.70	86.76
TPG	95.01	56.74	55.58	28.95	78.73
PromptSRC	94.24	48.52	59.58	24.66	72.72
FedOTP	94.25	76.86	72.50	33.88	85.91

The experimental results strongly demonstrate the effectiveness of the two key features of OTDPP. First, OTDPP adopts optimal transport, it works particularly effectively on datasets with significant visual domain shifts, and enhances model robustness. Second, the dual-prompt mechanism of OTDPP can effectively alleviate data heterogeneity, the global prompt captures transferable semantics, while the personalized component adapts to client-specific styles and category distributions. While FedOTP also employs a form of personalization, its performance, though strong, remains below that of OTDPP, suggesting that our integrated optimal transport and dual-prompt framework offers a more comprehensive solution. In contrast, competing methods either rely heavily on distillation and regularization for frozen CLIP representations (e.g., ProGrad and PromptSRC) or learn a single modality or a single prompt (e.g., FedPGP and TPG), and this makes competing methods more susceptible to data heterogeneity.

Although the algorithm in this paper demonstrates competitive overall performance, it is worth noting that its accuracy on the aircraft dataset is relatively low compared to other benchmarks. This is because the aircraft dataset involves fine-grained visual classification, where categories correspond to different model variants of aircraft. The distinctions between categories rely on subtle local discriminative features. Compared to datasets like Caltech101, where different classes exhibit large semantic gaps and clearly distinct features, the aircraft dataset is more challenging. It better reflects the algorithm's performance in few-shot fine-grained learning scenarios. This deeply reveals a limitation of OTDPP under its current design: for fine-grained classification tasks that heavily rely on subtle local discriminative features, its performance improvement encounters a bottleneck.

4.4. The Impact of Data Heterogeneity on Performance

To evaluate the robustness of the OTDPP algorithm to data heterogeneity, we adjust the alpha parameter of the Dirichlet distribution used for client data partitioning and set it to 0.1, 0.5, and 1.0. A smaller α value indicates a more severe non-IID characteristic of client data, while a larger α value means a more uniform data distribution among clients. The experimental results are shown in Table 2, where OTDPP achieves the best average performance (75.73%) at $\alpha = 0.1$ where data heterogeneity is most severe, and when α

increases to 0.5 and 1.0, the average accuracy across the five datasets decreases to 69.30% and 65.65%, respectively. This seemingly counterintuitive performance trend is determined by the design orientation of OTDPP for highly heterogeneous non-IID scenarios and the intrinsic characteristics of different datasets.

Table 2. Performance Evaluation under Various Data Heterogeneity.

	Caltech101	DTD	Eurosat	Aircraft	Flowers
0.1	95.01	79.26	78.75	38.88	86.76
0.5	92.50	74.73	61.25	31.25	86.76
1.0	91.50	64.36	60.00	26.25	86.15

The identical accuracy of the Flowers dataset at $\alpha = 0.1$ and $\alpha = 0.5$ is a reproducible normal phenomenon. The Oxford Flowers-102 dataset has distinct visual discriminative features between categories with low fine-grained recognition difficulty, and our experiment adopts a strict few-shot setting (eight labeled samples per category) for all datasets. Under this setting, the OTDPP model can stably capture the core category features of flowers regardless of the moderate changes in data heterogeneity (α = 0.1 and α = 0.5), and thus the classification accuracy remains unchanged. Even when α increases to 1.0, the accuracy only slightly drops to 86.15%, which further verifies that the Flowers dataset is less sensitive to data heterogeneity due to its simple task characteristics. All experimental results for this dataset are based on three independent runs with different random seeds, and the standard deviation is 0, which fully guarantees the reproducibility of the results.

The performance decline with the increase in α is the direct result of the adaptive characteristics of the OTDPP algorithm. OTDPP is designed to solve the problem of client drift in traditional federated learning under highly non-IID data by decoupling global shared prompts and local personalized prompts, and the personalized prompt module is the core to improve the model's adaptation to local data. When α is large, the global model trained on such data is already nearly optimal and has strong generalization ability for all clients, because the common knowledge across clients is sufficient to support accurate classification. At this time, the personalized prompt module loses its effective learning target and is prone to overfitting to the limited local few-shot samples. This overfitting makes the model deviate from the nearly optimal global model, and the personalized module even produces a negative effect instead of the expected adaptive effect, leading to a gradual decline in the overall performance. In particular, for datasets with large domain shifts and high fine-grained requirements, the performance decline is more significant with the increase in α, because these datasets rely more on the personalized adaptation of the model to local data characteristics in non-IID scenarios, while the homogeneous data scenario weakens the value of the personalized module.

In contrast, when $\alpha = 0.1$, the global model trained by traditional federated learning methods will suffer from severe performance degradation due to serious client drift. At this time, the dual-prompt mechanism of OTDPP can play its maximum role: the global prompt captures the limited common knowledge across clients, and the local personalized prompt accurately adapts to the unique data distribution of each client. The optimal transport strategy further enhances the fine-grained alignment of cross-modal features under heterogeneous data.

4.5. Ablation Study

To evaluate the effectiveness of the personalized method in the OTDPP algorithm, we replace the dual-prompt metehod with a simpler single-prompt method (named OT) and compare its performance with the original OTDPP algorithm. We keep other conditions

unchanged in the experiment, and the experimental results on five benchmark datasets are shown in Figure 3.

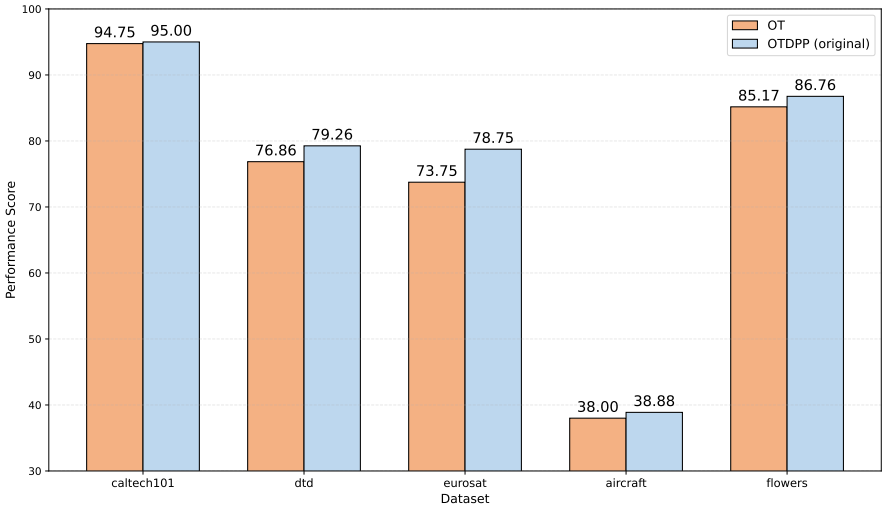

Figure 3. Performance improvements brought by key OTDPP modules.

The hierarchical prompt injection mechanism (OTDPP) outperforms the OT algorithm on all evaluated datasets, and this verifies the overall effectiveness of integrating the dual-prompt mechanism. Notably, OTDPP achieves better performance on Caltech101, DTD, EuroSAT, and Flowers.

The varying degrees of improvement achieved by the algorithm in this paper, compared to the single-prompt OT version, can be attributed to the characteristics of the datasets. For datasets like EuroSAT, which involve significant domain shifts, the dual-prompt mechanism performs well. The global prompt maintains cross-client semantic consistency, while the personalized prompt adapts to client-specific variations. In contrast, FGVC-Aircraft involves fine-grained classification. The basic OT alignment already achieves robust local feature matching, thereby reducing the marginal gains from personalization.

This paper further investigates the impact of sample size on the algorithm. During the training process, we adjusted the sample size within the range of [2, 4, 8], with corresponding results shown in Figure 4. The horizontal axis represents the sample size, while the vertical axis denotes the average test accuracy. It can be observed that as the sample size increases, the performance of all methods gradually improves.

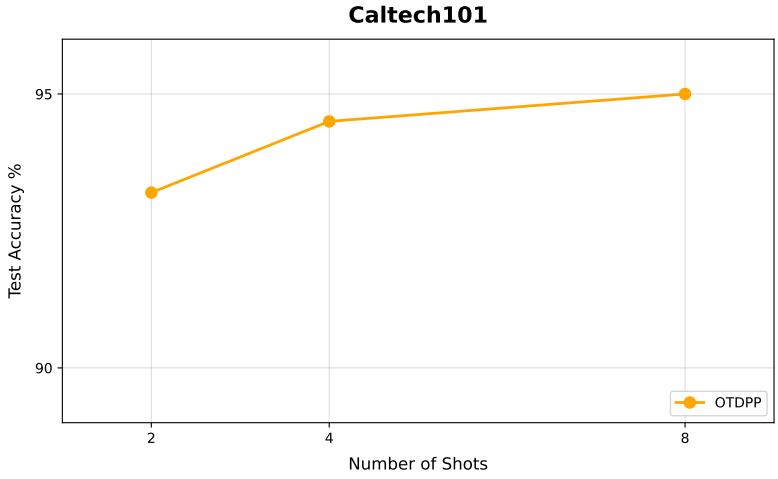

Figure 4. Performance with the different number of shots.

As shown in Figure 5, we demonstrate the impact of prompt depth in our proposed method on Caltech101. Overall, performance improves as the prompt depth increases.

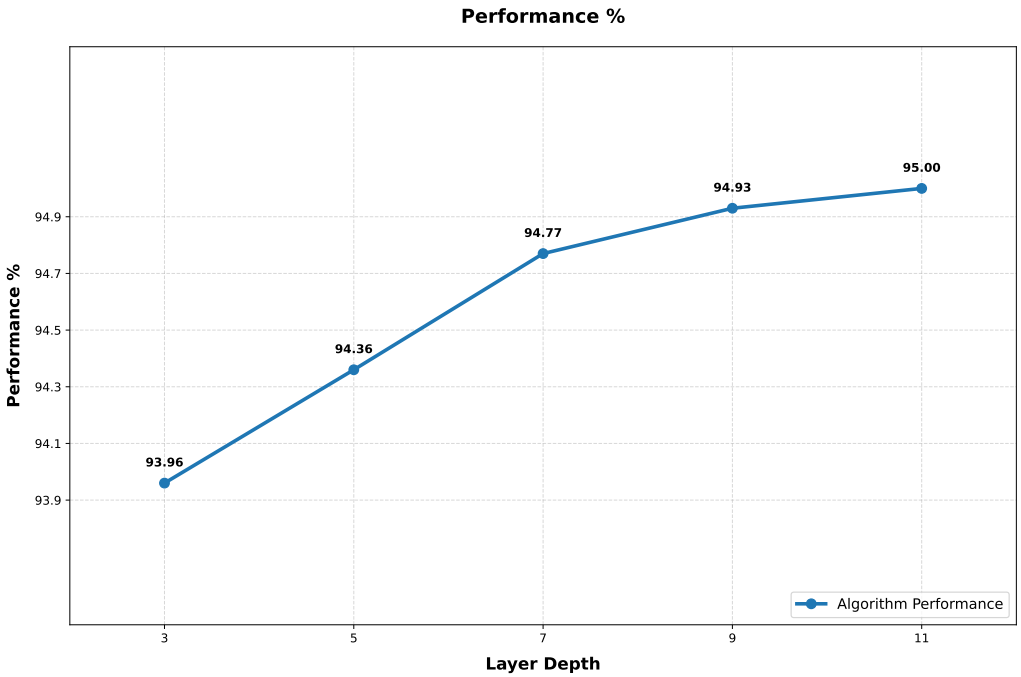

Figure 5. Ablation on prompt depth.

In summary, the ablation study confirms that the personalized federated learning module is a key contributor to the overall performance of the OTDPP framework. Removing this module causes the model's accuracy to decrease on all types of datasets, and this verifies the rationality of integrating personalized design into the prompt tuning paradigm. The experimental results clearly show that the full-version OTDPP algorithm has more robust and superior overall performance compared with the non-personalized version.

4.6. Computing and Communication Overhead Analysis

To quantitatively verify the computational and communication efficiency of the proposed method, we conducted detailed performance analysis under the typical configuration with a contextual size of (2, 4, 512). The total number of parameters in the introduced adapter module is 4096, and all are trainable parameters. Among them, 2048 parameters are designed as personalized parameters, and the remaining 2048 are global parameters. This design allows only 2048 parameters to be transmitted per round of parameter synchronization. Compared with the complete CLIP model as the backbone network, our method only introduces an additional 0.004752% of trainable parameters. This result shows that the proposed algorithm achieves efficient adaptation by updating a tiny fraction of parameters of the base model and thus ensures low communication and storage costs. This has critical advantages for federated learning and resource-constrained learning scenarios.

As shown in Figure 6, the experimental results demonstrate that OTDPP has significant advantages in terms of convergence speed and stability, fully reflecting the training efficiency and robustness of this method.

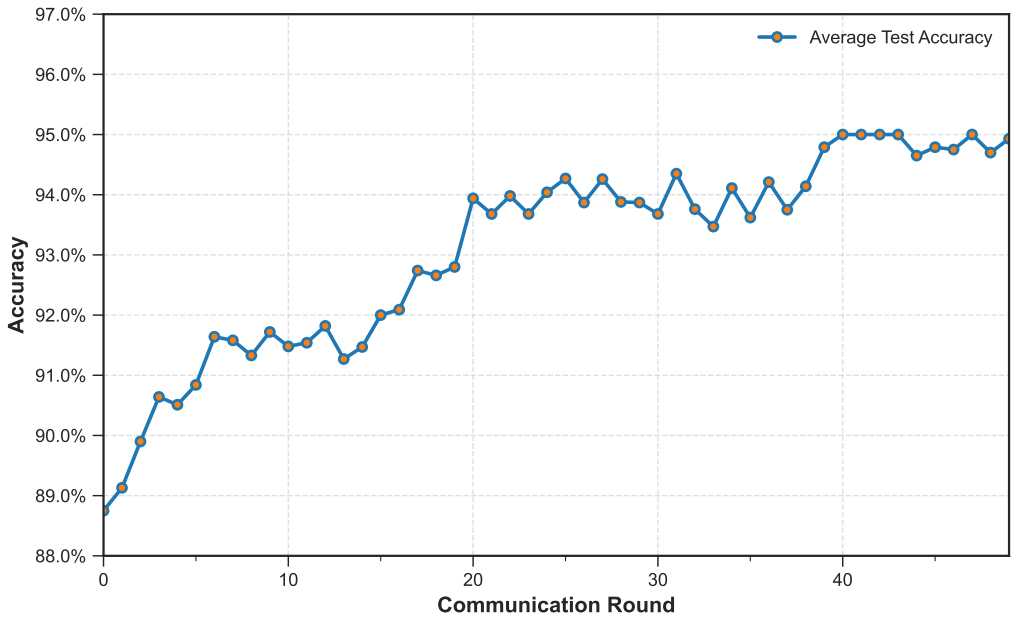

Figure 6. Accuracy learning curve on Caltech101.

5. Conclusions

This paper focuses on federated learning scenarios based on large-scale vision–language pre-trained models and systematically studies the challenge of personalized adaptation for cross-modal prompt learning under non-independent and identically distributed data. Existing methods struggle to balance global knowledge sharing and fine-grained alignment of local features when dealing with heterogeneous data distributions across clients, and this issue leads to degraded model performance.

The core contributions of the proposed algorithm lie in three aspects. First, it adopts a hierarchical prompt parameter decoupling design, which divides learnable prompt parameters into a global head and a local tail and injects them into multiple network layers. This design effectively alleviates the mismatch between the global model and local data distributions in traditional federated learning. Second, it applies optimal transport theory to compute the optimal transport distance between text prompt tokens and visual prompt tokens and achieves fine-grained alignment of cross-modal features. Third, it injects only a small number of prompt parameters into visual and text encoders and freezes large backbone networks. This design greatly reduces client-side computational load and communication overhead and enables continuous evolution of personalization capabilities while preserving privacy.

We conduct experimental validation of this work on existing public datasets. These datasets provide controllable benchmarks for comparison, and we recognize the importance of evaluating OTDPP in more realistic federated scenarios, such as highly heterogeneous device participation, asynchronous updates and stricter communication constraints. In addition, applying this framework to fields with inherent confidentiality requirements, such as medical imaging and finance, forms a key research direction. The design of OTDPP retains personalized prompts locally and aggregates only global components. It conforms to the data minimization principle in nature and provides a promising architectural foundation for such sensitive scenarios. We will rigorously test this method in these realistic and confidential environments in future work and fully evaluate its practical robustness and privacy–utility trade-off.

Although the proposed algorithm achieves promising progress in multimodal personalized federated learning, it still has several limitations for further exploration. This paper

fixes the ratio of global and local prompt parameters, and a more promising direction is to develop dynamically adjustable prompt structures and adaptively adjust the ratio or dimension of global and local parameters. Current optimal transport is mainly applied to the alignment of local image features and text prompts, and future work can extend this mechanism to richer cross-modal interactions, such as video-language and audio-language scenarios. In addition, the performance of this method is sensitive to hyperparameters, and we adopt unified settings across different datasets. Fine-tuning these hyperparameters for each client may bring further improvements. Alignment based on optimal transport may become unstable under extremely heterogeneous conditions, and researchers can consider integrating data augmentation techniques to alleviate this problem. Moreover, researchers can explore the integration of OTDPP with privacy-preserving technologies such as differential privacy and homomorphic encryption and provide stricter privacy security guarantees while maintaining model utility.

Author Contributions: Conceptualization, L.S. and Z.L.; methodology, L.S. and Z.L.; software, L.S. and Z.L.; validation, L.S., Z.L., X.D., Y.Z. and X.Y.; formal analysis, L.S. and Z.L.; investigation, L.S., Z.L., X.D., Y.Z. and X.Y.; resources, L.S. and Z.L.; data curation, L.S., Z.L., X.D., Y.Z. and X.Y.; writing—original draft preparation, L.S. and Z.L.; writing—review and editing, L.S., Z.L., X.D., Y.Z. and X.Y.; visualization, X.D., Y.Z. and X.Y.; supervision, L.S. and Z.L.; project administration, L.S. and Z.L.; funding acquisition, L.S. and Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the National Natural Science Foundation of China (No. 62576129), the Key Research Project of Natural Science for Universities in Anhui Province (No. 2024AH050182) and the Anhui Provincial Natural Science Foundation (No. 2308085MF212).

Data Availability Statement: The data presented in this study are available in public domain repositories. The Caltech-101 dataset is accessible at https://data.caltech.edu/records/mzrjq-6wc02 (accessed on 24 February 2026); the DTD can be retrieved from https://www.robots.ox.ac.uk/~vgg/data/dtd/ (accessed on 24 February 2026); the EuroSAT dataset is available at https://github.com/phelber/eurosat (accessed on 24 February 2026); the FGVC-Aircraft dataset is obtainable from https://www.robots.ox.ac.uk/~vgg/data/fgvc-aircraft/ (accessed on 24 February 2026); and the Oxford Flowers-102 dataset can be accessed at https://www.robots.ox.ac.uk/~vgg/data/flowers/102/ (accessed on 24 February 2026). These datasets were directly used in compliance with their respective open access policies without additional restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, C.; Jiang, H.; Chen, J.; Zhao, Y.; Fu, S.; Jing, F.; Guo, Y. An overview of machine unlearning. *High-Confid. Comput.* **2025**, *5*, 100254.
2. Sun, X.; Cai, Y.; Tao, Y.; Mai, L.; Huang, J.Z. LogoML: An Open Machine Learning Library for Distributed Big Data Analytics. *Big Data Min. Anal.* 2025, *accepted*.
3. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
4. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763.
5. Protection, F.D. *General Data Protection Regulation (GDPR)*; Intersoft Consulting: Hamburg, Germany, 2018; Volume 24.
6. Lialin, V.; Deshpande, V.; Rumshisky, A. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv* **2023**, arXiv:2303.15647. [CrossRef]
7. Wang, P.; Lu, W.; Lu, C.; Zhou, R.; Li, M.; Qin, L. Large Language Model for Medical Images: A Survey of Taxonomy, Systematic Review, and Future Trends. *Big Data Min. Anal.* **2025**, *8*, 496–517. [CrossRef]
8. Cai, Z.; Pang, J.; Li, Y.; Huang, Y.; Xie, Z. A Comprehensive Survey of Federated Open-World Learning. *IEEE Trans. Netw. Sci. Eng.* **2025**, *13*, 208–224. [CrossRef]

9. He, Z.; Wang, Y.; Cai, Z. Federated Continual Learning with Bounded Forgetting via Diffusion-Based Generative Replay in Edge Computing. *IEEE Trans. Mob. Comput.* **2025**, *25*, 3001–3017. [CrossRef]
10. Chen, X.; Pang, J.; Sun, T. Deep reinforcement learning based resource provisioning for federated edge learning. *High-Confid. Comput.* **2025**, *5*, 100264.
11. Chang, K.; Sun, H.; Wan, J.; Zhang, N.; Liu, Y.; Yang, K.; Shu, Z.; Xia, J.; Zhou, X. FedCE: A Contrast Enhancement Federated Learning Method for Heterogeneous Medical Named Entity Recognition. *Tsinghua Sci. Technol.* **2025**, *30*, 2384–2398. [CrossRef]
12. Tavassolian, F.; Abbasi, M.; Ramezani, A.; Taherkordi, A.; Khosravi, M.R. ResFed: An accurate and light federated multi-shot pre-trained model on edge devices. *Tsinghua Sci. Technol.* **2024**, *30*, 1539–1551. [CrossRef]
13. He, Z.; Li, Y.; Cai, Z. Personalized Federated Learning via Gradient-Fusion and Gradient-Decoupling for Heterogeneous Data. *IEEE Trans. Mob. Comput.* **2025**, *25*, 2956–2972. [CrossRef]
14. He, Z.; Wang, L.; Cai, Z. Clustered federated learning with adaptive local differential privacy on heterogeneous iot data. *IEEE Internet Things J.* **2023**, *11*, 137–146. [CrossRef]
15. Chang, Y.; Shi, X.; Zhao, X.; Chen, Z.; Ma, D. DP2FL: Dual Prompt Personalized Federated Learning in Foundation Models. *arXiv* **2025**, arXiv:2504.16357. [CrossRef]
16. Ghiasvand, S.; Alizadeh, M.; Pedarsani, R. pFedMMA: Personalized Federated Fine-Tuning with Multi-Modal Adapter for Vision-Language Models. *arXiv* **2025**, arXiv:2507.05394.
17. Luo, J.; Chen, C.; Wu, S. Mixture of Experts Made Personalized: Federated Prompt Learning for Vision-Language Models. *arXiv* **2024**, arXiv:2410.10114. [CrossRef]
18. Lu, X.; Shen, P.; Tsao, Y.; Kawai, H. Cross-modal Knowledge Transfer Learning as Graph Matching Based on Optimal Transport for ASR. *arXiv* **2025**, arXiv:2505.13079. [CrossRef]
19. Li, H.; Huang, W.; Wang, J.; Shi, Y. Global and local prompts cooperation via optimal transport for federated learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 12151–12161.
20. Chen, G.; Yao, W.; Song, X.; Li, X.; Rao, Y.; Zhang, K. Prompt Learning with Optimal Transport for Vision-Language Models. *arXiv* **2023**, arXiv:2210.01253. [CrossRef]
21. Zhu, X.; Zhu, B.; Wang, S.; Zhao, K.; Zhang, H. Enhancing CLIP Robustness via Cross-Modality Alignment. *arXiv* **2025**, arXiv:2510.24038. [CrossRef]
22. Khattak, M.U.; Rasheed, H.; Maaz, M.; Khan, S.; Khan, F.S. Maple: Multi-modal prompt learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19113–19122.
23. Khattak, M.U.; Naeem, M.F.; Naseer, M.; Van Gool, L.; Tombari, F. Learning to prompt with text only supervision for vision-language models. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 4230–4238.
24. Franklin, L.; Boonmee, A.; Wongsuwan, K. Vision-Driven Prompt Optimization for Large Language Models in Multimodal Generative Tasks. *arXiv* **2025**, arXiv:2501.02527. [CrossRef]
25. Zheng, H.; Yang, S.; He, Z.; Yang, J.; Huang, Z. Hierarchical cross-modal prompt learning for vision-language models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–20 October 2025; pp. 1891–1901.
26. Guo, T.; Guo, S.; Wang, J.; Tang, X.; Xu, W. Promptfl: Let federated participants cooperatively learn prompts instead of models–federated learning in age of foundation model. *IEEE Trans. Mob. Comput.* **2023**, *23*, 5179–5194.
27. Qiu, C.; Li, X.; Mummadi, C.K.; Ganesh, M.R.; Li, Z.; Peng, L.; Lin, W.Y. Text-driven prompt generation for vision-language models in federated learning. *arXiv* **2023**, arXiv:2310.06123.
28. Cui, T.; Li, H.; Wang, J.; Shi, Y. Harmonizing generalization and personalization in federated prompt learning. *arXiv* **2024**, arXiv:2405.09771. [CrossRef]
29. Zhu, B.; Niu, Y.; Han, Y.; Wu, Y.; Zhang, H. Prompt-aligned gradient for prompt tuning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 15659–15669.
30. Khattak, M.U.; Wasim, S.T.; Naseer, M.; Khan, S.; Yang, M.H.; Khan, F.S. Self-regulating prompts: Foundational model adaptation without forgetting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 15190–15200.