FDPT: Federated Discrete Prompt Tuning for Black-Box Visual-Language Models

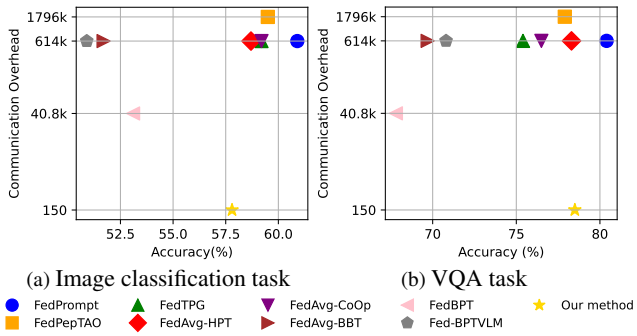
Anonymous submission

Abstract

*General-purpose Vision-Language Models (VLMs) have driven major advancements in multimodal AI. Fine-tuning these models with task-specific data enhances adaptability to various downstream tasks but suffers from privacy risks. While potential solutions like federated learning can address user data privacy concerns, model protection is also essential. Other methods that rely on black-box VLM APIs usually require the access of prediction logits, leaving them open to inversion attacks. Moreover, addressing the challenges of tuning complexity and data transmission efficiency in federated VLM scenarios is also crucial. To address these challenges, we propose FDPT—a **f**ederated **d**iscrete **p**rompt **t**uning method utilizing black-box VLM APIs. The black-box VLM is restricted to output text only, without relying on traditional prediction logits. In the client optimization stage, we design an AI agent-driven, token-level prompt optimization method that employs a Generate-Feedback mechanism to iteratively learn task experience. Additionally, we use an Exploration-Exploitation Balance strategy to drive a Progressive Prompt Refinement Chain-of-Thought, adaptively controlling the optimization scale. In the global aggregation stage, we propose a Semantic-similarity-guided Evolutionary Computation method, filtering local discrete tokens based on semantic similarity to enable unsupervised selection of representative tokens. Experimental results show that, compared to eight state-of-the-art methods, FDPT significantly improves accuracy in traditional image classification and visual question-answer tasks, reduces communication overhead, and produces highly transferable optimized prompts. Moreover, it demonstrates greater robustness to data heterogeneity.*

1. Introduction

General-purpose Vision-Language Models (VLMs) [44] integrate extensive visual and linguistic information, propelling advancements in generative AI content (AIGC) [35]. They can be fine-tuned to meet downstream task requirements [27]; however, this process introduces privacy risks [11, 37]. Federated Learning (FL) [22] addresses these concerns by decentralizing data storage and mitigating single-point data breaches, thereby preserving user privacy through the exchange of model weights or embeddings instead of raw data. Numerous approaches [17, 29, 38] leverage FL for fine-tuning large models; for instance, FedTPG [29] collaboratively generates context-aware prompts across clients, enhancing generalization for diverse classification tasks.



(a) Image classification task (b) VQA task

Figure 1. Comparison of accuracy and communication overhead across image classification and visual question-answering (VQA) tasks in federated settings on TinyImageNet and RealWorldQA datasets. FedAvg-BBT, FedBPT, and Fed-BPTVLM are configured as black-box VLM methods. Our proposed method *FDPT* achieves superior accuracy, even reaching white-box levels in VQA tasks, while significantly reducing communication overhead. For configurations and additional results, refer to Section 4.

However, safeguarding the privacy of VLMs themselves remains critical and requires further exploration. Current approaches typically utilize black-box VLMs [32, 40] to protect model parameters and apply zero-order optimization to estimate model gradients, enabling prompt optimization. For example, FedBPT [32] uses prompt exchange and gradient-free optimization, significantly reducing communication and memory costs by avoiding model parameter access and back-propagation on client devices. However, these approaches have several limitations.

- Existing black-box methods output prediction logits, which pose a risk of inversion attacks [5, 47] that can infer model parameters.
- These zero-order optimization strategies [32, 40] often lack precision, diminishing the accuracy and adaptability

of prompt fine-tuning.
- Soft prompt methods [14, 46] lack interpretability and generate substantial communication overhead in federated learning settings.

To address these challenges, we introduce FDPT, a novel federated and discrete prompt optimization method. This method utilizes a black-box VLM that outputs only text, without traditional prediction logits, ensuring complete privacy for large models. In the client-side optimization phase, we design an *Agent-based Client Prompt Optimization* for token-level prompts, leveraging a *Generate-Feedback Mechanism* to enable continuous task-specific learning. Additionally, the *Exploration-Exploitation Balance Strategy* guides the Agent's *Progressive Prompt Refinement Chain-of-Thought* to incrementally refine the scope of prompt optimization. In the global aggregation phase, inspired by human *Voting* activity, we propose a *Self-attention-guided Evolutionary Computation* method. This approach clusters local discrete tokens based on semantic similarity and selects representative tokens according to semantic-similarity weights, enabling efficient zero-order prompt optimization without additional data.

The contributions of this paper are as follows:
- We propose FDPT, a federated black-box discrete prompt optimization framework that ensures privacy for both users and VLMs.
- We introduce an AI agent-driven method for discrete, token-level clients' prompt optimization.
- We propose a semantic-similarity-guided evolutionary computation method that uses semantic similarity to unsupervisedly select representative tokens, achieving effective knowledge aggregation.
- Extensive experiments demonstrate that our method achieves superior performance across various text-vision tasks, exhibits low communication overhead, and is resilient to data heterogeneity.

2. Background & Related Work

Federated Learning. Federated Learning (FL) [22] is a decentralized approach allowing multiple clients to collaboratively train a shared model without sharing raw data, thus preserving privacy. Initially developed for privacy-preserving applications on mobile and edge devices [24], FL mitigates privacy risks by aggregating model updates instead of raw data. Recent works have further improved FL's robustness in handling data heterogeneity [26], label noise [15], and personalization needs [9]. For instance, recent studies[19, 49] introduce methods to address label noise and diverse data distributions in FL using public data alignment and noise-resistant loss functions, enhancing model robustness in real-world settings. Additionally, confidence-weighted aggregation in a variational expectation-maximization framework has been applied to manage client variability and domain-specific differences, achieving more personalized model outcomes [51]. Contrastive learning approaches [31, 50] are also used to align model representations across clients, effectively correcting local biases and improving accuracy, particularly in image classification tasks [20]. These advancements collectively address key challenges in FL, making it more adaptable and effective in distributed, real-world environments.

Prompt Tuning. Prompt tuning [52] is an emerging technique for adapting large pre-trained models to specific tasks with minimal additional training. In VLMs, prompt tuning [13] bridges visual and textual modalities, enabling tasks like image captioning [41, 43]. The Prompt tuning can be represented as finding an optimal prompt $\mathbf{p}^*$ to maximize the model's performance on the dataset $\mathcal{D} = \{(\mathbf{x}, \mathbf{y})\}$:

$$\mathbf{p}^* = \arg\max f_{\text{VLM}}(\mathbf{p}; \mathcal{D}) \tag{1}$$

where $f_{\text{VLM}}(\mathbf{p}; \mathcal{D})$ denotes the performance of the VLM using prompt $\mathbf{p}$ over the dataset $\mathcal{D}$. By introducing task-specific prompts, the VLMs can be adapted to new scenarios effectively and efficiently. For instance, recent research has applied prompt tuning in various ways: Adversarial Prompt Tuning (AdvPT) enhances robustness against adversarial attacks [45]; Attribute-Guided Prompt Tuning (ArGue) uses visual attributes from language models to improve classification accuracy [33]; and Distribution-Aware Prompt Tuning (DAPT) considers feature distributions for better generalization across domains [6]. These methods highlight the flexibility and adaptability of prompt tuning in VLMs, making it an essential tool for vision-language applications.

3. Method

3.1. Problem Formulation

The task of black-box discrete prompt optimization in Vision-Language Models (VLMs) is to find an optimal prompt $\mathbf{p}^*$ within a discrete search space $\mathcal{P}$, composed of a finite set of candidate prompts $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N\}$, relying solely on VLMs observable outputs such as logits ($\mathcal{O}_{\text{logits}}$), without access to model gradients. Thus, the objective is to maximize the model's performance on the centralized dataset $\mathcal{D} = \{(\mathbf{x}, \mathbf{y})\}$:

$$\mathbf{p}^* = \arg\max_{\mathbf{p} \in \mathcal{P}} f_{\text{VLM}}(\mathbf{p} \mid \mathcal{O}_{\text{logits}}; \mathcal{D}) \tag{2}$$

However, the above approach still presents privacy risks for both users and VLMs. To address this, we introduce federated learning into black-box VLM prompt tuning tasks, proposing a novel black-box federated discrete prompt tuning framework. Our framework offers privacy protection on multiple levels, *VLM Privacy*: Restricting black-box VLMs to output only text prevents inversion attacks. *Client*

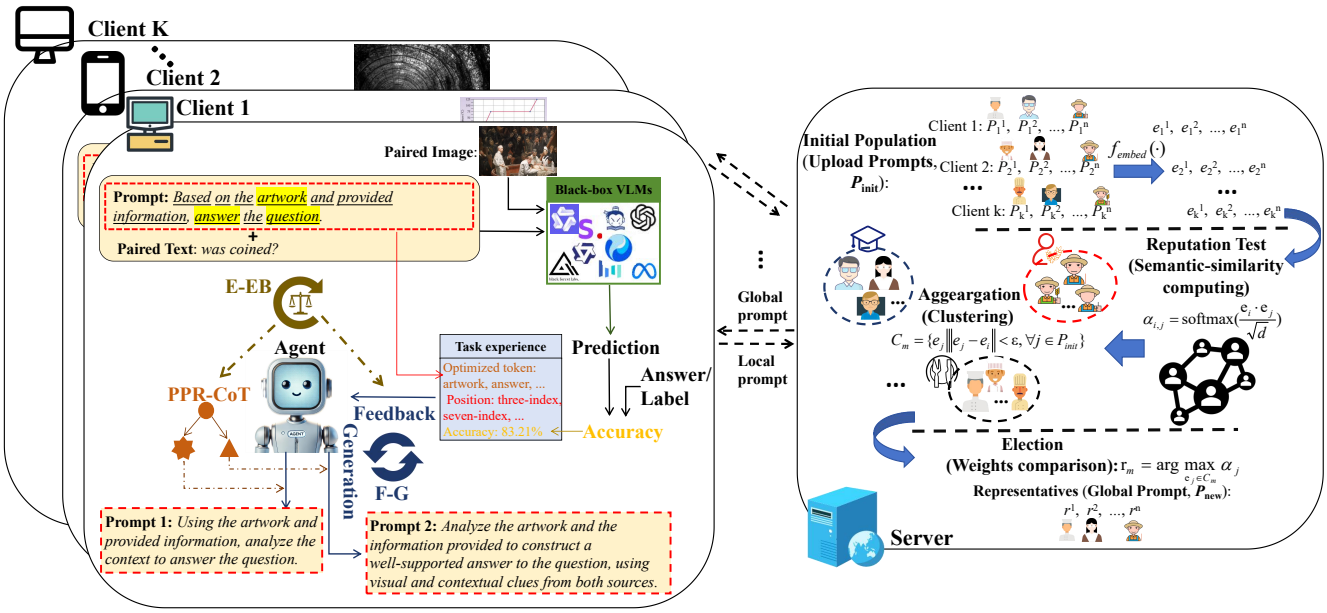


Figure 2. Overview of the *FDPT*. Each client utilizes an *Agent-based Client Prompt Optimization* method for token-level local prompt optimization. The optimized local prompt is then uploaded to the server, where a *Semantic-Similarity-Guided Evolutionary Computation* method is applied to select *Representatives* (*Global Prompt*) from the *Initial Population* (*Uploaded Prompts*). In the next round, the server broadcasts the *Global Prompt* to all clients. Additionally, the abbreviations in client module are explained in detail in Section 3.3.

Data Privacy: Data is maintained in a distributed manner, with tokens used to perform knowledge exchange. In this FL scenario, each client k maintains a private local dataset $\mathcal{D}_k = \{(\mathbf{x}_k, \mathbf{y}_k)\}$ and performs prompt optimization independently on its data. Clients locally optimize their prompts to produce $\mathbf{p}_k$ values, which are then aggregated by a central server to maximize overall VLM performance across all clients:

$$\mathbf{p}^* = \arg \max_{\mathcal{P}=\{\mathbf{p}_k\}_{k=1}^K} \sum_{k=1}^{K} f_{\text{VLM}}(\mathbf{p}_k \mid \mathcal{O}_{\text{text}}; \mathcal{D}_k) \quad (3)$$

where $\mathcal{O}_{\text{text}}$ represents the text-only output from the VLM, which prevents exposure of model internals. This federated approach thus enables secure and privacy-preserving prompt optimization in a collaborative, multi-client setting.

3.2. Design Overview

The algorithm overview is shown in Figure 2 . Following the conventional federated learning paradigm [22], our approach executes the following steps for each global communication round, *Initialization Broadcast:* the server distributes the *Global Prompt* to each client, *Client Optimization:* each client uses an *Agent-based Client Prompt Optimization* method to perform personalized token-level prompt optimization, *Server Aggregation:* each client uploads the locally optimized prompt to the server, where knowledge aggregation is performed using the *Semantic-similarity-guided Evolutionary Computation* method. The following sections provide a detailed explanation of each step in our approach.

3.3. Agent-based Client Prompt Optimization

To realize token-level prompt optimization, inspired by popular AI Agent strategies [16, 25], we design a *Prompt Optimization Agent* (PO Agent) on each client, as illustrated in Figure 2. In addition to traditional Agent operations such as data perception, reasoning, and execution, we propose three strategies: *Generation-Feedback Mechanism* (G-F), *Exploration-Exploitation Balance Strategy* (E-EB), and *Progressive Prompt Refinement Chain-of-Thought* (PPR-COT) to support the Agent in achieving effective prompt optimization. The optimization process of the PO Agent is shown in Algorithm 1. Below, we provide a detailed explanation of these methods.

Generation-Feedback Mechanism. Drawing inspiration from the forward and backward propagation process in deep learning [2], we propose a novel *Generation-Feedback Mechanism* (G-F) for Agents. This mechanism begins with forward propagation to obtain predictions, followed by feedback-driven adjustments: (1) The Agent generates an optimized prompt to guide the black-box VLM's predictions. (2) Each round's prompt optimization strategy and prediction results are stored as historical experience (i.e., feedback data) and fed back to the Agent in the next round. By incorporating this mechanism, the Agent leverages not only its inherent knowledge but also gains new learning pathways through task interaction. Unlike tradi-

Algorithm 1 Agent-Based Client Prompt Optimization for Black-Box VLMs

Input: Black-box VLM API with text predictions, initial prompt $\mathcal{P}$, accuracy threshold X, last iteration strategy $\mathcal{S}_{\text{last}}$, Accuracy $\mathcal{F}$
Output: Optimized prompt $\mathcal{P}_{\text{optimized}}$

for each optimization round **do**
  Store $\{\mathcal{S}_{\text{last}}, \mathcal{F}\}$ in history $\mathcal{H}$ ▷ G-F Mechanism
  At early iterations: ▷ E-EB Strategy
    E-EB calls *PPR-CoT.adjust_semantic*
  At later iterations:
    E-EB calls *PPR-CoT.refine_detail* with $\mathcal{H}$
end for
Return optimized prompt $\mathcal{P}_{\text{optimized}}$

procedure PPR-CoT.ADJUST_SEMANTIC
  Adjust semantic direction of $\mathcal{P}$
end procedure
procedure PPR-CoT.REFINE_DETAIL ▷ PPR-CoT
  Refine details of $\mathcal{P}$ using $\mathcal{H}$
end procedure

tional Agents that only execute reasoning [16, 25], this implicit task-based fine-tuning approach enables the Agent to better adapt to specific tasks, thereby realizing more effective prompt optimization.

Exploration-Exploitation Balance Strategy. Considering that task experience increases with each iteration, we propose a *Exploration-Exploitation Balance Strategy* (E-EB) to enable the Agent's adaptive use of commonsense knowledge alongside task-specific interaction experience. In the early stages of client prompt optimization, the Agent primarily relies on commonsense knowledge, forming optimization strategies based on new data. As optimization rounds continue and task experience grows, the Agent increasingly draws on historical task experience, enabling more tailored prompt optimization for client-specific data. This strategy is represented as $Strategy = \alpha \cdot Exploration + \beta \cdot Exploitation$, where α is the weight for exploration (new data) and $\beta = 1 - \alpha$ is the weight for exploitation (historical experience). A straightforward approach is to guide this balance strategy by a fitness (e.g., accuracy) threshold. For instance, when accuracy reaches $X\%$, the Agent's prompt template is updated to emphasize historical experience: "*Place greater emphasis on historical experience when making decisions.*" The adjustment formula for α is $\alpha = \max\left(0, 1 - \frac{\text{accuracy} - X}{100 - X}\right)$.

Progressive Prompt Refinement Chain-of-Thought. Existing chain-of-thought (CoT) strategies [12, 39] typically guide the Agent to make more effective decisions through refinement or step-by-step problem-solving. Given the characteristics of text generation tasks, we propose a *Progressive Prompt Refinement Chain-of-Thought* (PPR-COT), supported by the *Exploration-Exploitation Balance Strategy*. In initial stages of prompt optimization, the Agent conducts broader exploration based on client-specific data and commonsense knowledge, setting parameters such as semantic direction and sentence structure. In later stages, the Agent incrementally refines lexical and phrasing details based on accumulated historical task experience. This approach makes efficient use of client data by gradually narrowing the optimization scope to generate more suitable prompts. Compared to strategies [8, 32] focused solely on word-level adjustments, it better aligns with natural human writing patterns.

3.4. Semantic-similarity-guided Evolutionary Computation

After local optimization, each client uploads its optimized discrete prompt tokens to the server for knowledge exchange. We propose a *Semantic-similarity-Guided Evolutionary Computation* method to select representative tokens from the uploaded discrete tokens as the fused knowledge, without additional data support. This method uses *Reputation Test* (*Semantic-similarity computation*), *Aggregation* (*Clustering*), and *Election* (*Weight comparison*) to select representative tokens from the *Initial Population*, forming a set of *Representatives*. This process resembles a "*Voting*" activity [3] in human social dynamics: candidates are grouped into distinct *Communities* based on their characteristics, and an *Election* is conducted within each *Community* to identify *Representatives*. The algorithm steps are outlined in Algorithm 2, with further details provided below.

Initial Population. We treat the prompts uploaded by each client (denoted as $\mathcal{P}_{\text{client}}$) as the *Initial Population* $\mathcal{P}_{\text{init}}$:

$$\mathcal{P}_{\text{init}} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N\}. \tag{4}$$

We then use an encoder, such as the embedding layer of a LLM, to map tokens into a high-dimensional semantic space $\mathcal{E}$ as $\mathbf{e}_i = f_{\text{embed}}(\mathbf{t}_i) \in \mathcal{E}$, where $\mathbf{t}_i$ represents the i-th token, and f_{embed} is the embedding function producing a high-dimensional semantic embedding $\mathbf{e}_i$. In this space $\mathcal{E}$, semantically similar tokens are positioned closer together, reflecting their semantic relationships.

Reputation Test. We first evaluate the *reputation* of each token, defined as the semantic-similarity weight (denoted $\alpha_{i,j}$) that each token assigns to all other tokens. This is computed using a cosine-similarity mechanism [10] based on embeddings:

$$\alpha_{i,j} = \text{softmax}\left(\frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\sqrt{d}}\right), \tag{5}$$

where $\mathbf{e}_i$ and $\mathbf{e}_j$ represent the embeddings of tokens i and j, respectively, and d is the dimensionality of the embeddings. A higher semantic-similarity weight $\alpha_{i,j}$ signifies

a stronger influence over other tokens. We define the total semantic-similarity weight of token i as $\alpha_i = \sum_j \alpha_{i,j}$, which represents the token's semantic importance among all other tokens, referred to as its *Reputation*.

Aggregation. Next, tokens are clustered based on their semantic similarity by grouping embeddings that are close to each other in the high-dimensional semantic space:

$$\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_M\},\ \mathcal{C}_m = \{\mathbf{e}_j \,|\, \|\mathbf{e}_j - \mathbf{e}_i\| < \epsilon, \forall j \in \mathcal{P}_{\text{init}}\}, \tag{6}$$

where ϵ is a pre-defined distance threshold. $\mathcal{C}_M$ denotes a cluster, or *Community*. *Aggregation* provides the foundation for the subsequent *Election* phase, where representative tokens should be selected from clusters with distinct semantic characteristics to satisfy sentence syntax requirements. Tokens with higher weights have stronger semantic dependencies on other tokens and are more likely to be selected as *Representatives*.

Election. Finally, within each cluster (*Community*), the representative token is determined by semantic-similarity weights, or *reputation*:

$$\mathbf{r}_m = \arg \max_{\mathbf{e}_j \in \mathcal{C}_m} \alpha_j, \tag{7}$$

where $\mathbf{r}_m$ denotes the representative token in *Community* $\mathcal{C}_m$. Importantly, communities containing only a single token are discarded, as these tokens contain only client-specific information, following the principle of "majority rule" [30] in *Voting*. The selected tokens are then arranged in their original order (using, for example, a bubble sort [1]) to form a representative prompt, or the "*Representatives*", denoted as $\mathcal{P}_{\text{new}} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_M\}$.

Algorithm 2 Semantic-similarity-Guided Evolutionary Computation for Discrete Prompt Optimization

Input: K clients with discrete prompt tokens $\{\mathcal{P}_{\text{client}_k}\}_{k=1}^{K}$
Output: Global prompt tokens $\mathcal{P}_{\text{new}}$

Initialization *Initial population* $\mathcal{P}_{\text{init}} = \{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$
for each token $\mathbf{t}_i \in \mathcal{P}_{\text{init}}$ **do**
  Forward $\mathbf{t}_i$ to embeddings $\mathbf{e}_i = f_{\text{embed}}(\mathbf{t}_i)$
end for
for each token $\mathbf{t}_i$ **do** ▷ *Reputation Test*
  for each token $\mathbf{t}_j$ **do**
    Semantic-similarity weight according to Eq.5
  end for
  Get total semantic-similarity weight $\alpha_i = \sum_j \alpha_{i,j}$
end for
Cluster tokens according to Eq.6 ▷ *Aggregation*
for each *Community* $\mathcal{C}_m \in \mathcal{C}$ **do** ▷ *Election*
  Select representative token according to Eq.7
  if $\mathcal{C}_m$ contains only one token **then**
    Discard *Community* $\mathcal{C}_m$
  end if
end for
Form *Representatives* $\mathcal{P}_{\text{new}} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_M\}$ from selected representative tokens

4. Evaluation

4.1. Evaluation Setup

Backbone API. In our experiments, we selected four backbone models: Qwen2-VL-72B [34], InternVL2-Llama3-76B [1], TeleMM [23], and GPT4o [2].

Dataset and Evaluation. We conduct experiments on two tasks: traditional image classification and Visual Question Answer (VQA). For image classification, we use the MNIST [7], CIFAR-10, CIFAR-100, and TinyImageNet (TinyImg) [18] datasets. For VQA, we utilize the Art, Chemistry, and Finance sections of the MMMU dataset [42], along with the RealWorldQA (RWQA) dataset [3]. The task objective is to leverage optimized prompts for accurate image classification on the image datasets, and for VQA, to assist in parsing textual content within the datasets to enable accurate classification or decision-making.

We use *Accuracy* as the evaluation metric for the image classification task. To maintain consistency across tasks, we also constrain predictions in the VQA task to the range of available labels and evaluate performance using *Accuracy*.

Comparison Baselines. We compare our method against eight state-of-the-art (SOTA) approaches: FedPrompt [46], FedPepTAO [4], FedTPG [29], HPT [36], CoOp [48], BBT [8], FedBPT [32], BPTVLM [40], and a manual prompting method (Manual Prompt). Notably, HPT [36], CoOp [48], BBT[8], and BPTVLM [40] are not originally designed for federated learning. To adapt these methods to a federated setting, we apply a weighted averaging aggregation approach, following the implementations of FedPrompt [46]and FedBPT [32]. The results are reported using federated versions of these models: FedAvg-HPT, FedAvg-CoOp, FedAvg-BBT, and Fed-BPTVLM. Additionally, we categorize these methods into *White-box* and *Black-box* approaches based on whether they require access to model parameters. The *White-box* methods include FedPrompt [46], FedPepTAO [4], FedTPG [29], FedAvg-HPT, and FedAvg-CoOp, while the *Black-box* methods consist of Manual, FedAvg-BBT, FedBPT [32], and Fed-BPTVLM. Detailed descriptions of all baselines are provided in Appendix A.

Implementation & Hyperparameters . The federated learning setup for our experiments is as follows: we use a default of 10 clients, with a client participation rate of

[1] https://huggingface.co/OpenGVLab/InternVL2-Llama3-76B
[2] https://platform.openai.com/docs/models/gpt-4o
[3] https://huggingface.co/datasets/xai-org/RealworldQA

Table 1. Performance Comparison on CV Classification and VQA Tasks. The accuracy results are measured in percentages (%). **Bold values** indicate the overall best results, and underlined values indicate the best results among *Black Box* methods. Under all experimental conditions, our method approaches the performance of *White Box* algorithms and significantly outperforms *Black Box* methods.

Backbone		Method	CV Classification				VQA				Average
			MNIST	Cifar-10	Cifar-100	TinyImg	Art	Chemistry	Finance	RWQA	
Qwen2-VL-72B	White Box	FedPrompt	90.4	**79.5**	**73.5**	**60.9**	74.6	**63.3**	**70.5**	**80.4**	**74.1**
		FedPepTAO	86.7	78.3	72.9	59.5	69.1	58.8	68.5	77.9	71.5
		FedTPG	84.2	76.8	71.3	59.2	72.4	57.5	62.6	75.4	69.9
		FedAvg-HPT	86.9	78.4	70.2	58.7	**75.7**	53.5	65.9	78.3	71.0
		FedAvg-CoOp	85.7	78.1	71.4	59.2	70.2	61.8	63.3	76.5	70.8
	Black Box	Manual	70.7	62.2	53.5	47.3	54.7	45.8	51.8	58.3	55.5
		FedAvg-BBT	78.9	72.3	62.7	51.7	63.5	53.9	60.3	69.7	64.1
		FedBPT	80.0	72.0	64.2	53.1	68.0	50.6	59.0	67.8	64.3
		Fed-BPTVLM	80.7	73.1	62.3	50.9	65.2	53.3	63.3	70.8	64.9
		Our method	**90.5**	78.5	71.8	57.8	74.0	60.8	67.2	78.5	72.4
InternVL2-Llama3-76B	White Box	FedPrompt	88.1	**81.3**	**74.4**	**60.2**	70.2	**58.8**	62.9	**76.1**	**71.3**
		FedPepTAO	87.3	77.8	71.2	58.5	65.2	55.3	63.0	75.4	69.2
		FedTPG	86.2	79.3	71.7	57.4	68.0	54.2	60.7	73.8	68.9
		FedAvg-HPT	85.1	78.7	72.3	58.4	69.1	53.0	62.0	74.3	69.1
		FedAvg-CoOp	85.7	79.1	71.6	58.1	**71.8**	56.2	**63.9**	71.7	69.8
	Black Box	Manual	71.8	63.7	51.8	45.4	52.5	46.3	53.8	61.3	55.8
		FedAvg-BBT	80.0	71.1	65.2	52.7	60.2	51.9	53.8	70.5	63.2
		FedBPT	80.7	72.7	66.6	52.3	61.3	50.6	58.4	67.1	63.7
		Fed-BPTVLM	79.5	71.4	67.2	53.0	64.1	53.0	56.7	69.8	64.3
		Our method	**89.6**	79.0	71.5	57.3	70.7	55.9	63.6	75.5	70.5
TeleMM	Black Box	Manual	69.4	65.5	53.3	45.8	53.0	45.0	52.5	57.9	55.3
		FedAvg-BBT	80.9	70.6	63.1	49.1	56.9	48.8	56.4	64.1	61.2
		FedBPT	80.0	71.3	61.9	50.9	60.8	51.5	57.4	66.3	62.5
		Fed-BPTVLM	79.1	71.0	62.7	49.7	59.1	52.3	57.7	68.5	62.5
		Our method	**88.4**	**78.9**	**69.3**	**57.4**	**64.6**	**55.7**	**61.6**	**75.0**	**68.9**
GPT4o	Black Box	Manual	74.4	68.2	59.1	52.8	61.9	58.6	61.6	71.9	63.6
		FedAvg-BBT	81.0	76.7	67.6	56.1	64.6	57.5	63.6	69.2	67.0
		FedBPT	83.2	78.3	68.4	56.6	69.1	59.3	68.5	69.9	69.2
		Fed-BPTVLM	81.7	79.0	67.9	57.3	67.4	57.9	65.2	71.6	68.5
		Our method	**92.5**	**85.3**	**73.9**	**62.7**	**76.8**	**66.7**	**72.8**	**78.5**	**76.1**

100% in each training round. Additionally, we adopt a *k-shot mechanism* for prompt tuning, following the approach in work [32], with the default value of k set to 50. For image classification datasets, we use the original training and test sets. For each Visual QA dataset, we randomly select 50 samples as the training set (from the validation set for MMMU and the test set for RealWorldQA), with the remaining samples used for testing. In addition, all experiments are conducted on two NVIDIA GeForce RTX A800 GPUs. Detailed hyperparameter settings are provided in Appendix B.

4.2. Effectiveness Results

We first evaluate the effectiveness of our method on traditional CV classification tasks and VQA tasks. Experimental results show that, compared to other methods, our method demonstrates significant performance improvements in both tasks, particularly excelling in black-box environments, with only a small gap from white-box algorithms.

Traditional CV Classification Tasks. In traditional CV classification tasks, our method achieves outstanding performance across all model configurations. For example, in the GPT4o black-box setting, our method achieves improvements of 11.2%, 7.0%, 5.5%, and 6.1% over the closest-performing FedBPT [32] method on the MNIST, Cifar-10, Cifar-100, and TinyImg datasets, respectively. In the TeleMM black-box setting, our method also performs exceptionally well, achieving 88.4% and 78.9% accuracy on MNIST and Cifar-10, respectively, which is an improvement of 9.3% and 8.3% over FedAvg-BBT. Additionally, for the Qwen2-VL-72B model, our method achieves accuracies of 78.5% and 71.8% on the Cifar-10 and Cifar-100

datasets, respectively, representing improvements of 7.1% and 5.3% over FedBPT [32], further demonstrating its advantage in traditional CV tasks.

VQA Tasks. Our method also shows strong performance in VQA tasks. In the GPT4o black-box setting, our method achieves improvements of 11.1%, 7.4%, 4.3%, and 8.6% over the FedBPT [32] method on the Art, Chemistry, Finance, and RWQA datasets, respectively. In TeleMM setting, our method's performance remains very stable, achieving improvements of 7.7% and 10.9% over FedAvg-BBT on the Art and RWQA datasets, respectively.

4.3. Prompt Transferability Evaluation

Table 2. Transferability Experiment Results. The accuracy results are measured in percentages (%). The prompts optimized by our method maintain excellent performance when transferred to other backbone models.

Model	Cifar-10	Cifar-100	TinyImg
Qwen2-VL→ TeleMM	80	70.5	58.1
InternVL2→ GPT4o	87.8	72.3	63.5

In the prompt transferability experiment, we evaluate the performance of prompts optimized on one backbone model when applied to another, as shown in Table 2. The results indicate that prompts exhibit a degree of transferability, maintaining high robustness across models with minimal accuracy loss, as referenced in Table 1. Additionally, compared to the Qwen2-VL transfer results, the InternVL2 transfer showed better accuracy, further validating the high applicability and robustness of prompts between certain models.

4.4. Communication Overhead Experiment

Table 3. Communication Overhead Experiment results. The measurements are in *Tensors* within the *PyTorch* Framework. **Bold values** indicate the best results, and underlined values indicate the second-best results. Our method achieves a significant reduction in communication overhead.

Type	Method	Parameters
White Box	FedPrompt	614k
	FedPepTAO	1796k
	FedTPG	614k
	FedAvg-HPT	614k
	FedAvg-CoOp	614k
Black Box	FedAvg-BBT	614k
	FedBPT	40.8k
	Fed-BPTVLM	614k
Our method		**150**

In the communication overhead experiment, our method shows a significant advantage in parameter efficiency, as illutrated in Table 3. Specifically, white-box methods like FedPrompt [46] and FedTPG [29] and most black-box methods such as FedAvg-BBT and Fed-BPTVLM require hundreds of thousands of parameters, whereas our method requires only 150. This substantial reduction effectively minimizes latency and resource consumption associated with communication. This advantage primarily arises from a key difference in algorithm design: while other methods use soft prompts for knowledge exchange between clients, our method employs discrete prompt tokens. Soft prompts typically require a large number of continuous parameters, whereas discrete prompt tokens enable knowledge transfer with only a small set of parameters, significantly reducing communication overhead.

4.5. Ablation Studies

Table 4. The Improvements Ablation Studies Results. The accuracy is measured in percentages (%). Performance decreases with the removal or modification of any key component, underscoring the effectiveness of the improvements. "w Embedding-fedavg" denotes the use of a weighted averaging aggregation method, as in FedBPT [32] and FedAvg-BBT. "w Random search" refers to the removal of *weight comparison*, using random search instead to select *representative tokens*. "w/o Clustering Strategy" indicates the removal of the *clustering* step. "w Fixed Clusters" implies a fixed number of clusters, which affects the count of *representative tokens*. Additionally, the abbreviations in client-site are explained in detail in Section 3.3.

Site	Method	Cifar-10	Cifar-100	TinyImg
Server	w Embedding-fedavg	71.9	62.5	54.2
	w Random search	65.3	57.3	53.9
	w/o Clustering Strategy	76.5	70.6	55.9
	w Fixed Clusters	77.8	69.3	56.3
Client	w/o G-F Mechanism	71.8	65.5	53.5
	w/o E-EB Strategy	76.9	70.8	56.2
	w/o PPR-CoT	75.3	69.3	56.7
Our method		78.5	71.8	57.8

In the improvements ablation study based on Qwen2-VL-72B API, our method consistently achieves the best performance across all metrics, further validating the effectiveness of its key components, as shown in Table 4. Compared to cases where specific modules are removed or replaced, our method demonstrates significant improvements on the Cifar-10, Cifar-100, and TinyImg datasets. **Server-side experiments.** Our method improves accuracy on the Cifar-10 and Cifar-100 datasets by 2.0% and 1.2%, respectively, compared to the setting without the clustering strategy. Furthermore, compared to the fixed clustering approach, our method achieves accuracy gains of 0.7% on Cifar-10 and 2.5% on Cifar-100, underscoring the importance of a dynamic clustering strategy. **Client-side experiments.** Removing the G-F mechanism results in a significant accuracy drop across all datasets. Furthermore, our method achieves

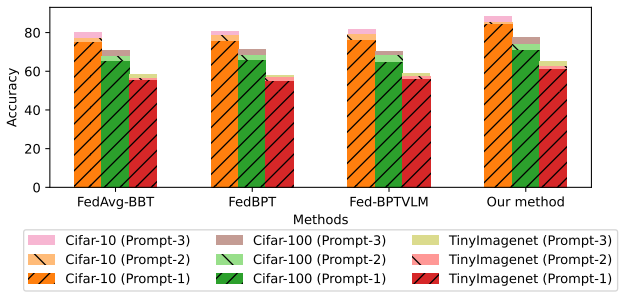


Figure 3. Initial Prompt Impact. *Prompt-1* represents a "shorter initial prompt," *Prompt-2* a "standard initial prompt," and *Prompt-3* a "detailed initial prompt." Across all datasets, our method consistently achieves the best performance under varying initial prompt settings.

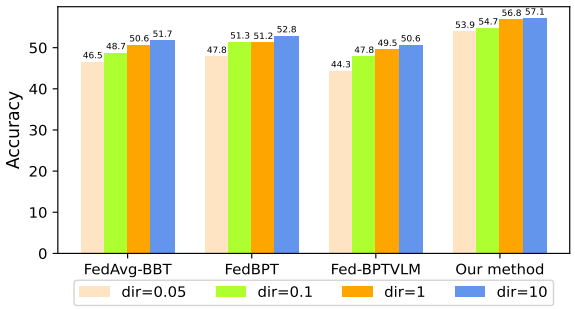


Figure 4. Heterogeneous Data Robustness Experiment results on the TinyImageNet dataset. The parameter "dir" is the *Dirichlet Distribution* hyperparameter; smaller values of dir indicate higher data heterogeneity across the 10 clients. Our method consistently achieves the best performance across varying heterogeneity settings, demonstrating strong robustness to data heterogeneity.

a 1.0% accuracy improvement on Cifar-100 and 1.6% on TinyImg compared to the configuration without the E-EB strategy. Compared to the configuration without the PPR-CoT strategy, our method improves accuracy by 3.2% on Cifar-10 and 2.5% on Cifar-100, underscoring the pivotal role of the PPR-CoT strategy in enhancing model performance.

Initial Prompt Impact. We evaluate the impact of the *Initial Prompt* on performance using GPT-4o as the backbone black-box VLM for classification tasks across datasets, comparing it with three typical black-box algorithms. Figure 3 shows that different prompt configurations significantly affect model performance, with our algorithm consistently achieving the best outcomes. Specifically, detailed prompts enhance model accuracy by providing richer information, which supports more effective learning and adaptation. In contrast, shorter prompts reduce accuracy, indicating that limited information constrains performance. Thus, prompt design is critical in classification tasks, and our algorithm exhibits strong robustness and performance across various prompt configurations. We present all *Initial Prompts* in Appendix C.

4.6. Heterogeneous Data Robustness Experiment

In the heterogeneous data robustness experiment, we use the Qwen2-VL-72B API as the backbone VLM model to evaluate the performance of different algorithms on heterogeneous data in the TinyImagenet dataset, as shown in Figure 4. Data is distributed across 10 clients using a *Dirichlet Distribution* [21] with varying hyperparameters *dir*, where a smaller Dirichlet parameter (e.g., dir=0.05) indicates greater data heterogeneity. Details of these heterogeneous data distributions are provided in Appendix D. The results demonstrate that all methods show some robustness to data heterogeneity, aligning with expectations [28, 32] for model stability on heterogeneous data. Additionally, as heterogeneity increases, our method exhibits superior performance, with minimal accuracy reduction, consistently outperforming other methods under all heterogeneity conditions. This highlights its notable advantage and robustness in handling highly heterogeneous data.

5. Conclusion

We propose FDPT, a federated black-box discrete prompt tuning framework based on VLMs to safeguard the privacy of both users and VLMs. FDPT uses an *Agent-based Client Prompt Optimization* strategy to achieve token-level prompt optimization on the client side without requiring VLMs to output traditional prediction logits, thereby preventing inversion attacks that could infer model parameters. It employs a *Semantic-similarity-guided Evolutionary Computation* method to perform unsupervised representative token selection. Moreover, discrete prompt optimization enhances interpretability, while discrete token exchange between clients offers high security and low communication overhead. In the extensive experiments, compared to eight state-of-the-art methods and a manual prompt approach, our method demonstrates superior accuracy in traditional computer vision classification and VQA tasks, lower communication overhead, transferable optimized prompts, and increased robustness to heterogeneous data.

References

[1] Owen Astrachan. Bubble sort: an archaeological algorithmic analysis. *ACM Sigcse Bulletin*, 35(1):1–5, 2003. 5

[2] Sebastian Banert, Jevgenija Rudzusika, Ozan Öktem, and Jonas Adler. Accelerated forward-backward optimization using deep learning. *SIAM J. Optim.*, 34(2):1236–1263, 2024. 3

[3] Marina F Barnea and Shalom H Schwartz. Values and voting. *Political psychology*, 19(1):17–40, 1998. 4

[4] Tianshi Che, Ji Liu, Yang Zhou, Jiaxiang Ren, Jiwen Zhou, Victor S Sheng, Huaiyu Dai, and Dejing Dou. Federated learning of large language models with parameter-efficient prompt tuning and adaptive optimization. *arXiv preprint arXiv:2310.15080*, 2023. 5

[5] Yiyi Chen, Heather Lent, and Johannes Bjerva. Text embedding inversion security for multilingual language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7808–7827, 2024. 1

[6] Eulrang Cho, Jooyeon Kim, and Hyunwoo J. Kim. Distribution-aware prompt tuning for vision-language models, 2023. 2

[7] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012. 5

[8] Shizhe Diao, Zhichao Huang, Ruijia Xu, Xuechun Li, Yong Lin, Xiao Zhou, and Tong Zhang. Black-box prompt learning for pre-trained language models. *Trans. Mach. Learn. Res.*, 2023, 2023. 4, 5

[9] Xiuwen Fang and Mang Ye. Robust federated learning with noisy and heterogeneous clients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10072–10081, 2022. 2

[10] Cong Gao, Wenfeng Li, Lijun He, and Lingchong Zhong. A distance and cosine similarity-based fitness evaluation mechanism for large-scale many-objective optimization. *Engineering Applications of Artificial Intelligence*, 133:108127, 2024. 4

[11] Aditya Golatkar, Alessandro Achille, Yu-Xiang Wang, Aaron Roth, Michael Kearns, and Stefano Soatto. Mixed differential privacy in computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8376–8386, 2022. 1

[12] Xu Gu, Xiaoliang Chen, Peng Lu, Zonggen Li, Yajun Du, and Xianyong Li. Agcvt-prompt for sentiment classification: Automatically generating chain of thought and verbalizer in prompt learning. *Eng. Appl. Artif. Intell.*, 132:107907, 2024. 4

[13] Xiang Gu, Shuchao Pang, Anan Du, Yifei Wang, Jixiang Miao, and Jorge Díez. Dynamic multimodal prompt tuning: Boost few-shot learning with vlm-guided point cloud models. In *ECAI*, pages 761–768. IOS Press, 2024. 2

[14] 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*, 2023. 2

[15] Xinyuan Ji, Zhaowei Zhu, Wei Xi, Olga Gadyatskaya, Zilong Song, Yong Cai, and Yang Liu. Fedfixer: Mitigating heterogeneous label noise in federated learning. In *AAAI*, pages 12830–12838. AAAI Press, 2024. 2

[16] Doha Kim and Hayeon Song. Designing an age-friendly conversational AI agent for mobile banking: the effects of voice modality and lip movement. *Int. J. Hum. Comput. Stud.*, 187: 103262, 2024. 3, 4

[17] Weirui Kuang, Bingchen Qian, Zitao Li, Daoyuan Chen, Dawei Gao, Xuchen Pan, Yuexiang Xie, Yaliang Li, Bolin Ding, and Jingren Zhou. Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5260–5271, 2024. 1

[18] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 5

[19] Jichang Li, Guanbin Li, Hui Cheng, Zicheng Liao, and Yizhou Yu. Feddiv: Collaborative noise filtering for federated learning with noisy labels. In *AAAI*, pages 3118–3126. AAAI Press, 2024. 2

[20] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10713–10722, 2021. 2

[21] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pages 965–978. IEEE, 2022. 8

[22] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020. 1, 2, 3

[23] Xiang Li, Yiqun Yao, Xin Jiang, Xuezhi Fang, Chao Wang, Xinzhang Liu, Zihan Wang, Yu Zhao, Xin Wang, Yuyao Huang, Shuangyong Song, Yongxiang Li, Zheng Zhang, Bo Zhao, Aixin Sun, Yequan Wang, Zhongjiang He, Zhongyuan Wang, Xuelong Li, and Tiejun Huang. Tele-flm technical report, 2024. 5

[24] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. Federated learning in mobile edge networks: A comprehensive survey. *IEEE communications surveys & tutorials*, 22(3):2031–2063, 2020. 2

[25] Haoxuan Ma, Yifan Liu, Qinhua Jiang, Brian Yueshuai He, Xishun Liao, and Jiaqi Ma. Mobility AI agents and networks. *IEEE Trans. Intell. Veh.*, 9(7):5124–5129, 2024. 3, 4

[26] Mulei Ma, Chenyu Gong, Liekang Zeng, Yang Yang, and Liantao Wu. Flocoff: Data heterogeneity resilient federated learning with communication-efficient edge offloading. *IEEE Journal on Selected Areas in Communications*, 42(11): 3262–3277, 2024. 2

[27] Yifei Ming and Yixuan Li. How does fine-tuning impact out-of-distribution detection for vision-language models? *International Journal of Computer Vision*, 132(2):596–609, 2024. 1

[28] J Nguyen, J Wang, K Malik, M Sanjabi, and M Rabbat. Where to begin? on the impact of pre-training and initialization in federated learning. arxiv 2022. *arXiv preprint arXiv:2210.08090*. 8

[29] 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. 1, 5, 7

[30] Mathias Risse. Arguing for majority rule*. *Journal of Political Philosophy*, 12(1), 2004. 5

[31] Guanxiong Shen, Junqing Zhang, Xuyu Wang, and Shiwen Mao. Federated radio frequency fingerprint identification powered by unsupervised contrastive learning. *IEEE Trans. Inf. Forensics Secur.*, 19:9204–9215, 2024. 2

[32] Jingwei Sun, Ziyue Xu, Hongxu Yin, Dong Yang, Daguang Xu, Yiran Chen, and Holger R Roth. Fedbpt: Efficient federated black-box prompt tuning for large language models. *arXiv preprint arXiv:2310.01467*, 2023. 1, 4, 5, 6, 7, 8

[33] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Argue: Attribute-guided prompt tuning for vision-language models, 2024. 2

[34] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 5

[35] Tao Wang, Yushu Zhang, Shuren Qi, Ruoyu Zhao, Xia Zhihua, and Jian Weng. Security and privacy on generative data in aigc: A survey. *ACM Computing Surveys*, 2023. 1

[36] Yubin Wang, Xinyang Jiang, De Cheng, Dongsheng Li, and Cairong Zhao. Learning hierarchical prompt with structured linguistic knowledge for vision-language models. In *AAAI*, pages 5749–5757. AAAI Press, 2024. 5

[37] Yubo Wang, Chaohu Liu, Yanqiu Qu, Haoyu Cao, Deqiang Jiang, and Linli Xu. Break the visual perception: Adversarial attacks targeting encoded visual tokens of large vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1072–1081, 2024. 1

[38] Feijie Wu, Zitao Li, Yaliang Li, Bolin Ding, and Jing Gao. Fedbiot: Llm local fine-tuning in federated learning without full model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3345–3355, 2024. 1

[39] Guang Yang, Yu Zhou, Xiang Chen, Xiangyu Zhang, Terry Yue Zhuo, and Taolue Chen. Chain-of-thought in neural code generation: From and for lightweight language models. *IEEE Trans. Software Eng.*, 50(9):2437–2457, 2024. 4

[40] Lang Yu, Qin Chen, Jiaju Lin, and Liang He. Black-box prompt tuning for vision-language model as a service. In *IJCAI*, pages 1686–1694, 2023. 1, 5

[41] Bowen Yuan, Sisi You, and Bing-Kun Bao. Self-pt: Adaptive self-prompt tuning for low-resource visual question answering. In *ACM Multimedia*, pages 5089–5098. ACM, 2023. 2

[42] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 5

[43] Jingjing Zhang, Shancheng Fang, Zhendong Mao, Zhiwei Zhang, and Yongdong Zhang. Fine-tuning with multi-modal entity prompts for news image captioning. In *ACM Multimedia*, pages 4365–4373. ACM, 2022. 2

[44] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[45] Jiaming Zhang, Xingjun Ma, Xin Wang, Lingyu Qiu, Jiaqi Wang, Yu-Gang Jiang, and Jitao Sang. Adversarial prompt tuning for vision-language models, 2024. 2

[46] Haodong Zhao, Wei Du, Fangqi Li, Peixuan Li, and Gongshen Liu. Fedprompt: Communication-efficient and privacy-preserving prompt tuning in federated learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2, 5, 7

[47] Xuejun Zhao, Wencan Zhang, Xiaokui Xiao, and Brian Lim. Exploiting explanations for model inversion attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 682–692, 2021. 1

[48] 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. 5

[49] Xiaochen Zhou and Xudong Wang. Federated label-noise learning with local diversity product regularization. In *AAAI*, pages 17141–17149. AAAI Press, 2024. 2

[50] Xiaokang Zhou, Qiuyue Yang, Xuzhe Zheng, Wei Liang, Kevin I-Kai Wang, Jianhua Ma, Yi Pan, and Qun Jin. Personalized federated learning with model-contrastive learning for multi-modal user modeling in human-centric metaverse. *IEEE J. Sel. Areas Commun.*, 42(4):817–831, 2024. 2

[51] Junyi Zhu, Xingchen Ma, and Matthew B Blaschko. Confidence-aware personalized federated learning via variational expectation maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24542–24551, 2023. 2

[52] Yu Zhu, Kang Li, Lequan Yu, and Pheng-Ann Heng. Memory-efficient prompt tuning for incremental histopathology classification. In *AAAI*, pages 7802–7810. AAAI Press, 2024. 2