A. Detailed Descriptions of All Baselines

Below are the implementation details of all baselines:

FedPrompt: In this method, only the soft prompt parameters (represented as framed pink nodes) require updating, aggregation (on the server), and redistribution.

FedPepTAO: Within each round, a random subset of devices is sampled. The prompt parameters from the previous round are transmitted to each selected device, where local updates are performed using the Adam optimizer. Afterward, each selected device returns the accumulated differences in the prompt parameters to the server.

FedTPG: The original FedTPG encodes the class names of the dataset and exchanges the encoder parameters between the server and clients for knowledge exchange. *However, this approach is incompatible with our prompt-based method.* To ensure fairness, we retain the computation of class names locally and utilize additional soft prompts for communication between the server and clients, aligning the conditions with other baselines. In this study, *we report the results of the improved FedTPG*. Additionally, we provide the results of the original FedTPG for reference. Specifically, using the ”Qwen2-VL-72B” model, the accuracy on MNIST, CIFAR-10, CIFAR-100, and TinyImageNet datasets is 85.9%, 75.5%, 73.6%, and 58.5%, respectively. For the Art, Chemistry, Finance, and RWQA tasks, the accuracy is 74.5%, 58.3%, 61.4%, and 74.3%, respectively. Furthermore, due to the use of model parameter communication between the server and clients, the original FedTPG incurs significant communication overhead, nearly *four times* that of the results presented in this paper.

FedAvg-HPT: *The original HPT does not inherently support a federated architecture*. To ensure a fair comparison, we extend HPT by integrating a federated learning (FL) framework. Specifically, we introduce an additional ”*initial prompt*” input into the text encoder and incorporate a soft prompt communication mechanism between the server and clients to optimize the prompt. During the server aggregation phase, we adopt the common weighted averaging strategy to achieve knowledge aggregation. Furthermore, in each global communication round, two local optimization iterations are performed: one to optimize the soft prompt and another to execute the original HPT optimization process.

Moreover, the improved methods for **FedAvg-CoOp**, **FedAvg-BBT**, **Fed-BPTVLM**, and **FedAvg-HPT** follow the above approach.

Manual: This method uses only the ”*initial prompt*” for inference. Its results serve as a fundamental benchmark to evaluate the optimization effectiveness of all approaches. The settings of this baseline follow on the *BBT* work.

FedBPT: Clients in FedBPT utilize a gradient-free optimization method (CMA-ES) to search for the optimal prompt distribution based on local data. This approach does not require clients to access the parameters of the pretrained language model (PLM); only the PLM's inference is performed during the search. The server aggregates the locally uploaded distributions to compute the globally optimal prompt distribution, which is then sent back to clients for the next round of optimization.

B. Detailed Experimental Setup

We provide a detailed experimental setup below.

Devices	**CPU**: Intel(R) Xeon(R) Gold 6348
	SSD: 100GB
	GPU: A800-80GB * 2
Software Tools	**CUDA** 12.1
	PyTorch 2.3.0
	Python 3.10
K-shot Training	***k***: 50
Datasets Setting	**All image classification datasets**: Original setting of train and test sets
	MMMU: train set includes 50 images in original validation set.
	test set includes the remaining samples.
	RealWorldQA: train set includes 50 images in original test set.
	test set includes the remaining samples.

Table 5. *Detailed experimental setup utilized in this paper.*

C. Various Initial Prompt Categories

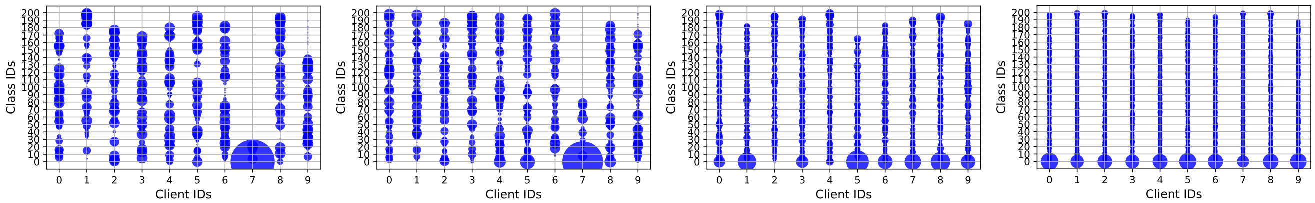
For each dataset, we provide three types of initial prompts to guide task understanding and execution:

- ***Prompt-1***: Represents a concise and minimalistic initial prompt for quick and straightforward interpretation.
- ***Prompt-2***: Represents a standard initial prompt with a balanced level of detail to ensure clarity.
- ***Prompt-3***: Represents a comprehensive and detailed initial prompt designed to provide thorough guidance and contextual information for complex tasks.
- **MNIST**
  - Classify the digit in the image.
  - Identify the handwritten digit shown in the image.
  - Please examine the image of the handwritten digit and determine its numeric value from 0 to 9, based on the shape and appearance.
- **Cifar-10**
  - Classify the object in the image.
  - Identify the category of the object shown in the image, such as an animal or a vehicle.
  - Please analyze the image content and classify the object into one of the categories, considering its visual features and context to determine if it's an animal, vehicle, or another object type.
- **Cifar-100**
  - Classify the object in the image.
  - Identify the category of the object shown in the image, considering its visual characteristics.
  - Please analyze the object in the image and determine the most suitable category based on its visual features and context.

- **TinyImagenet**
  - Classify the object in the image.
  - Identify the category of the object shown in the image based on its visual features.
  - Please analyze the object in the image and classify it according to its visual characteristics and context, selecting the most accurate category.
- **MMMU@Art**
  - Based on the artwork and provided information, answer the question.
  - Using the details of the artwork and the accompanying text, select the most accurate answer to the question.
  - Carefully analyze the artwork and related information provided in the text. Draw from the visual and contextual cues to choose the answer that best aligns with the question about this piece of art.
- **MMMU@Chemistry**
  - Based on the chemical information and image provided, answer the question.
  - Using the details in the chemical diagram and the accompanying text, select the most accurate answer to the question.
  - Carefully analyze the chemical structure shown in the image along with the information provided in the text. Use both visual and contextual clues to choose the answer that best addresses the chemistry-related question.
- **MMMU@Finance**
  - Based on the financial information and chart provided, answer the question.
  - Using the financial data and accompanying text, identify the most accurate answer to the question.
  - Carefully examine the financial chart and the provided information. Use both visual data and contextual insights to select the answer that best addresses the finance-related question.
- **RealworldQA**
  - Identify the correct answer based on the image and question.
  - Using the image and question provided, choose the correct answer from the options.
  - Carefully examine the image and interpret the question to determine the correct answer. Select the option that best fits, and respond with only the letter of the chosen answer.

D. Various Data Heterogeneity Settings

We present the data distribution from Figure 4 in the main text, which illustrates the heterogeneous data distribution of the TinyImageNet dataset across 10 clients under different Dirichlet hyperparameters, as shown below.



(a) dir = 0.05 (b) dir = 0.1 (c) dir = 1 (d) dir = 10

Figure 4. The sample distributions for all clients on the TinyImagenet datasets under the data heterogeneity settings with varying parameters dir. The size of each circle indicates the number of samples.