Vision Language Models-based Prompt Tuning for Federated Learning

DMQA Open Seminar (2026. 06. 05)

Data Mining & Quality Analytics Lab.

김다빈

발표자 소개

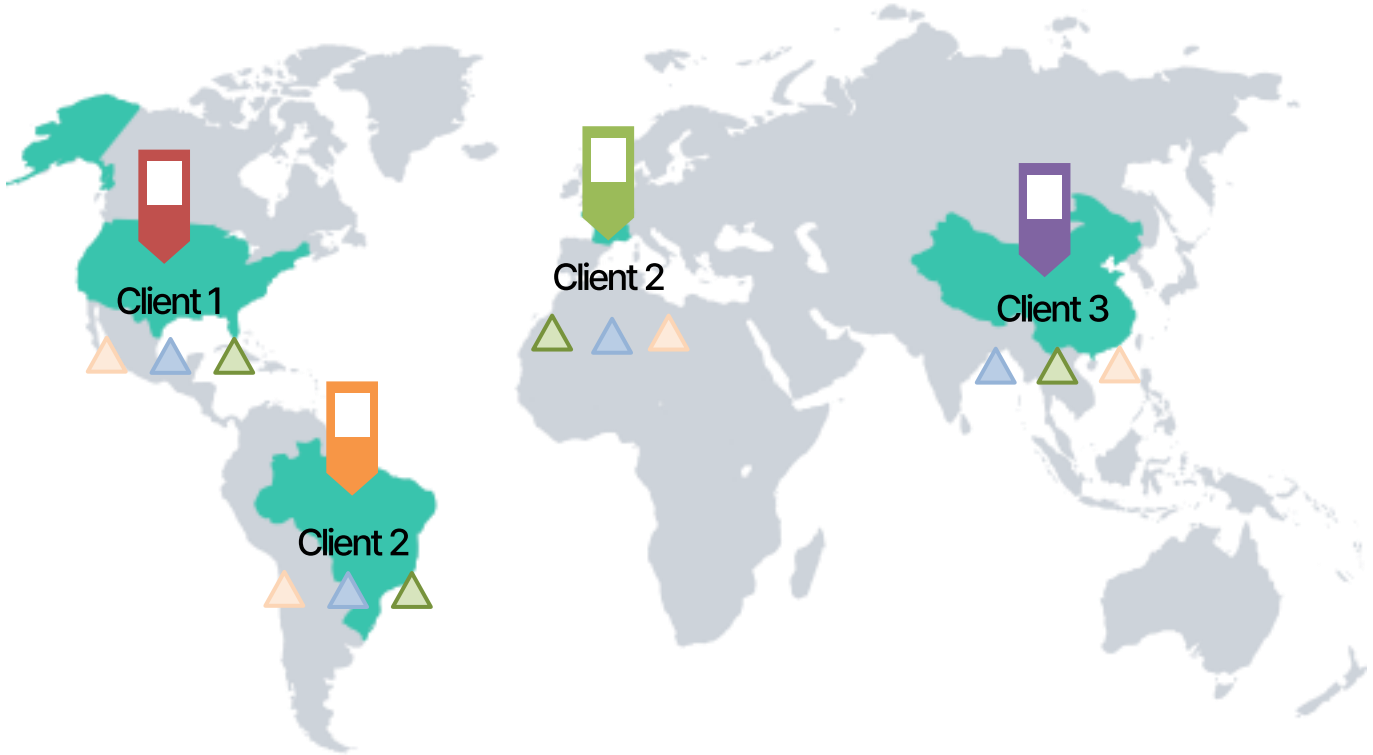
- **김다빈 (Dabin Kim)**
  - 고려대학교 산업경영공학과 대학원 재학
  - Data Mining & Quality Analytics Lab. (김성범 교수님)
  - M.S Student (2025.09 ~ Present)
- **Research Interest**
  - Federated Learning
  - Vision-Language Models
  - AI agent
- **Contact**
  - heydabins@korea.ac.kr

Introduction

Vision Language Models-based Prompt Tuning for Federated Learning

❖ Federated Learning

- 데이터 프라이버시를 보호하기 위해 각 클라이언트들이 자신의 데이터를 공유하지 않고,
  로컬에서 학습한 모델 파라미터만을 중앙 서버와 공유하여 협력적으로 글로벌 모델을 만드는 프레임워크

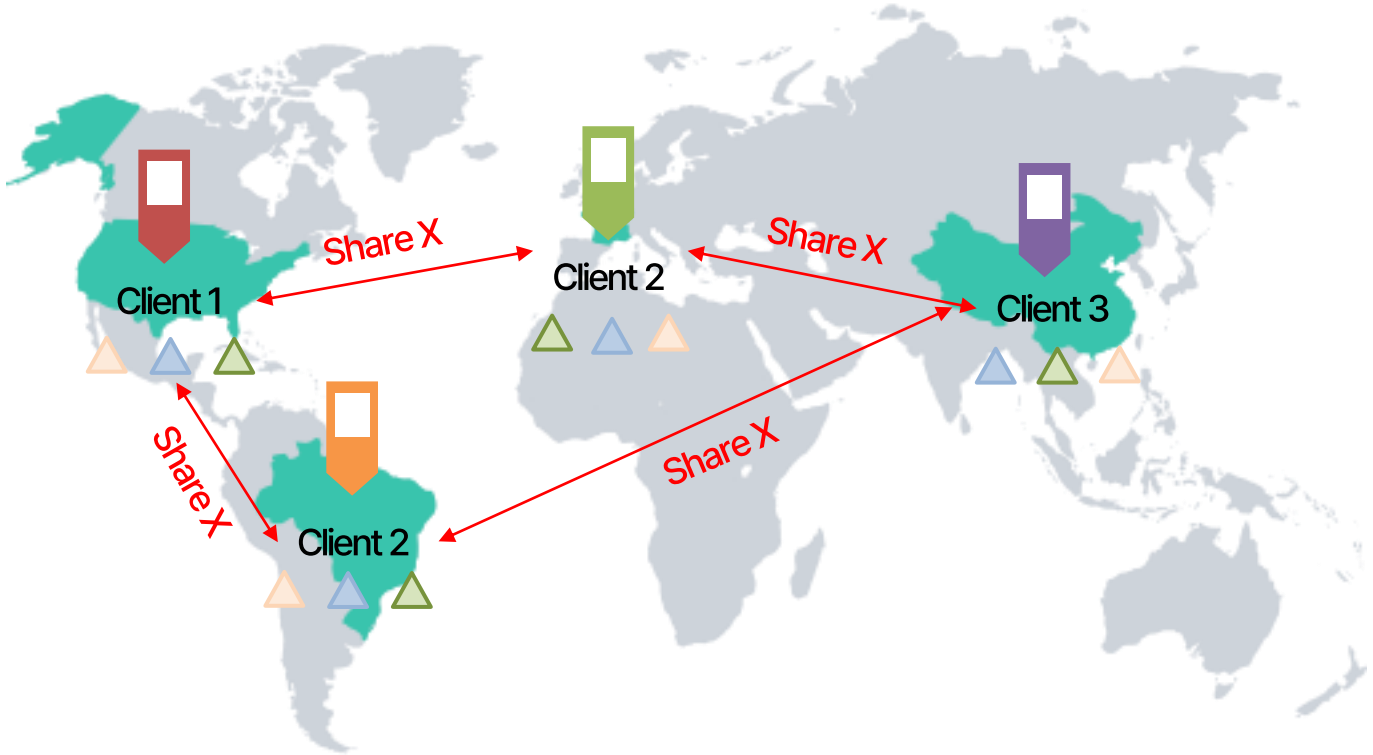


Introduction

Vision Language Models-based Prompt Tuning for Federated Learning

❖ Federated Learning

- 데이터 프라이버시를 보호하기 위해 각 클라이언트들이 자신의 데이터를 공유하지 않고,
  로컬에서 학습한 모델 파라미터만을 중앙 서버와 공유하여 협력적으로 글로벌 모델을 만드는 프레임워크

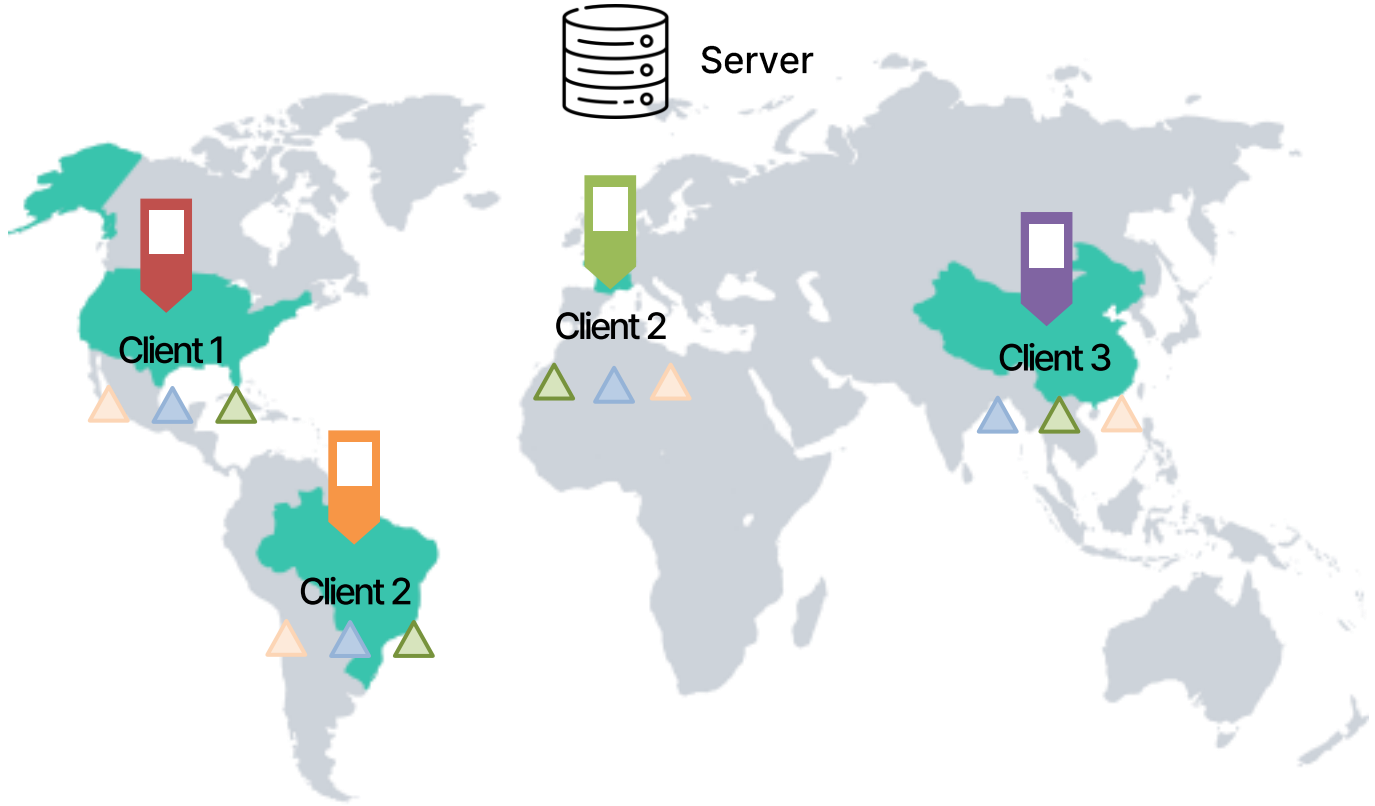


Introduction

Vision Language Models-based Prompt Tuning for Federated Learning

❖ Federated Learning

- 데이터 프라이버시를 보호하기 위해 각 클라이언트들이 자신의 데이터를 공유하지 않고,
  로컬에서 학습한 모델 파라미터만을 중앙 서버와 공유하여 협력적으로 글로벌 모델을 만드는 프레임워크

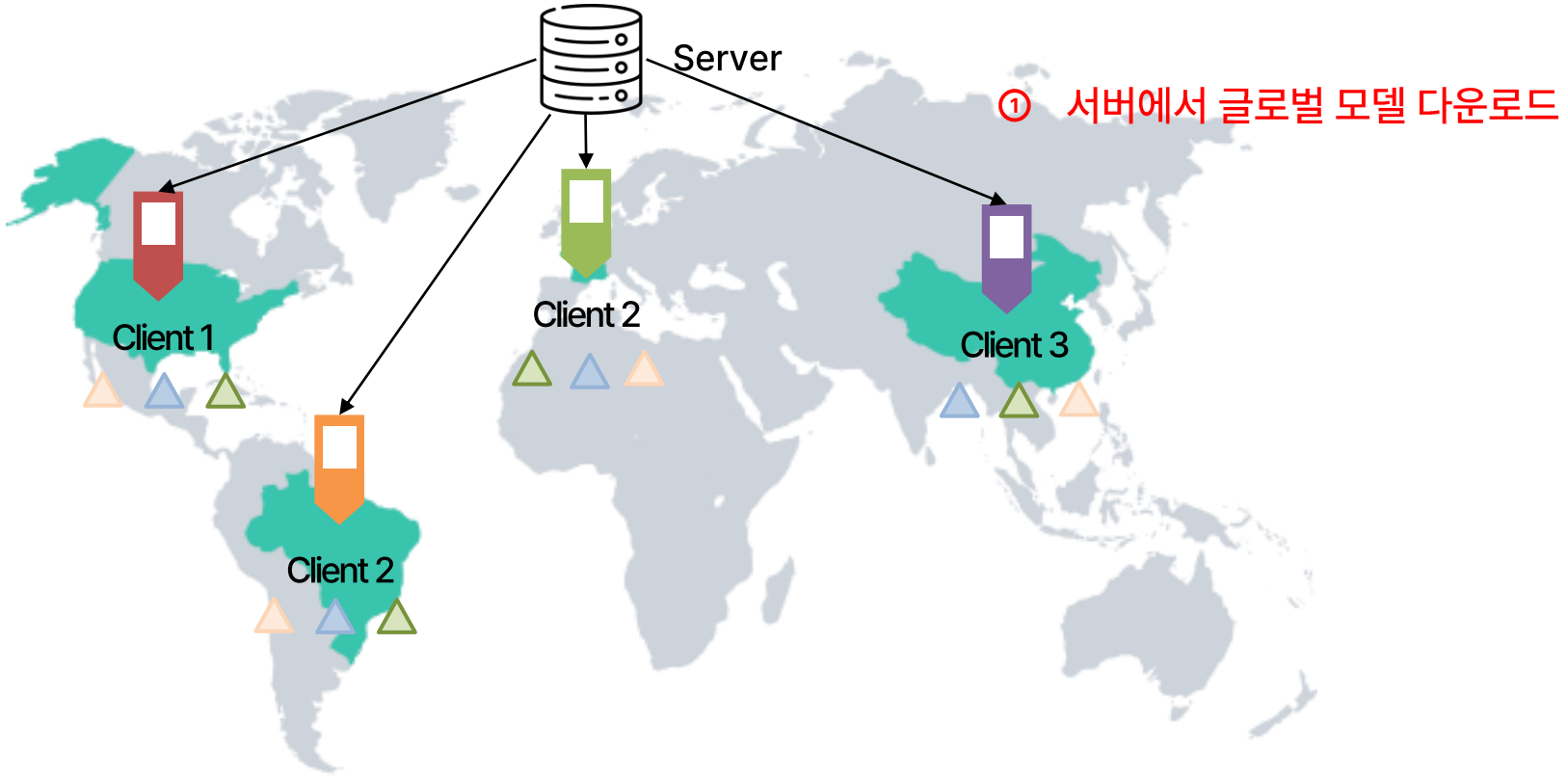


Introduction

Vision Language Models-based Prompt Tuning for Federated Learning

❖ Federated Learning

- 데이터 프라이버시를 보호하기 위해 각 클라이언트들이 자신의 데이터를 공유하지 않고,
  로컬에서 학습한 모델 파라미터만을 중앙 서버와 공유하여 협력적으로 글로벌 모델을 만드는 프레임워크

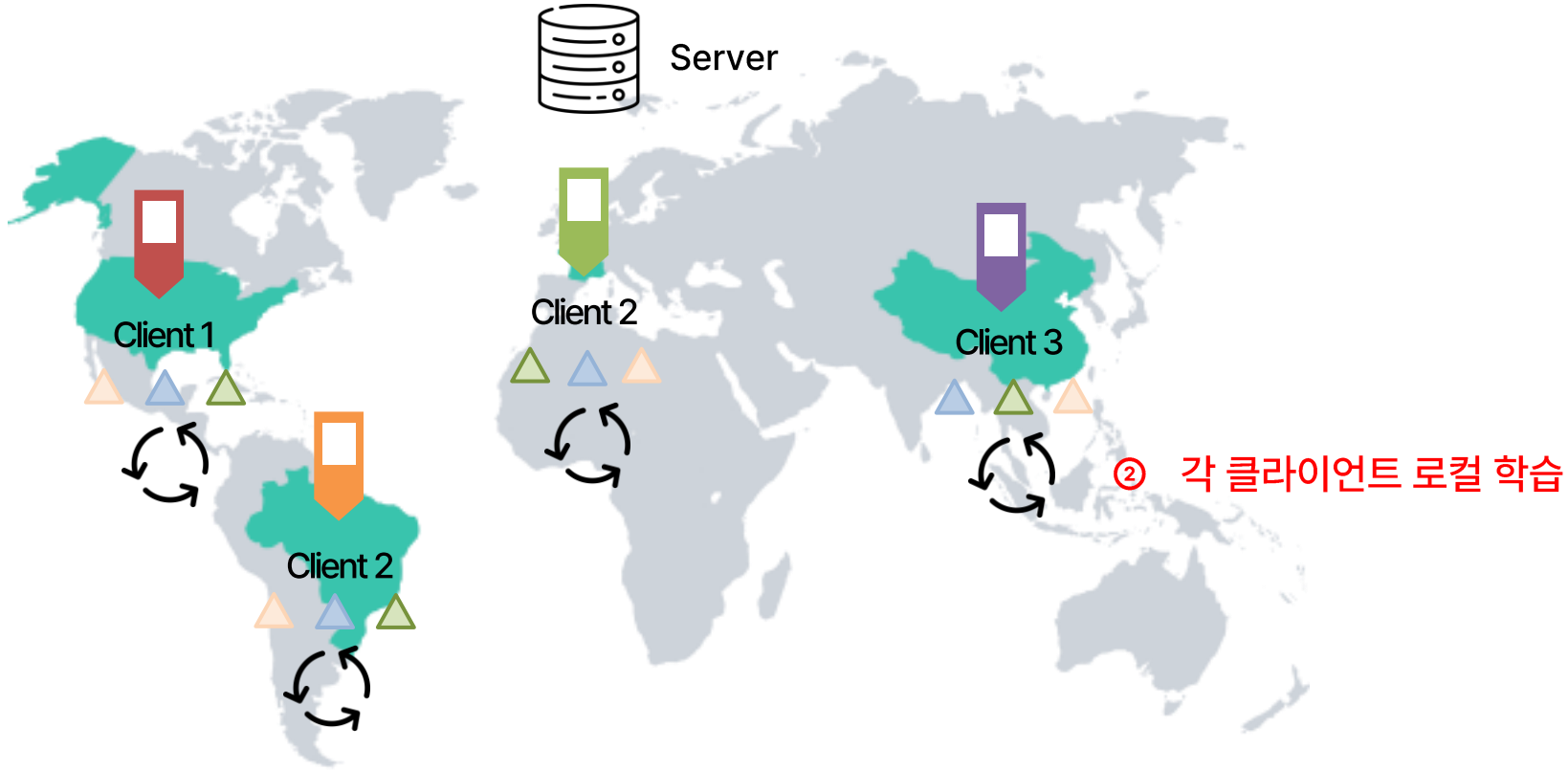


Introduction

Vision Language Models-based Prompt Tuning for Federated Learning

❖ Federated Learning

- 데이터 프라이버시를 보호하기 위해 각 클라이언트들이 자신의 데이터를 공유하지 않고,
  로컬에서 학습한 모델 파라미터만을 중앙 서버와 공유하여 협력적으로 글로벌 모델을 만드는 프레임워크

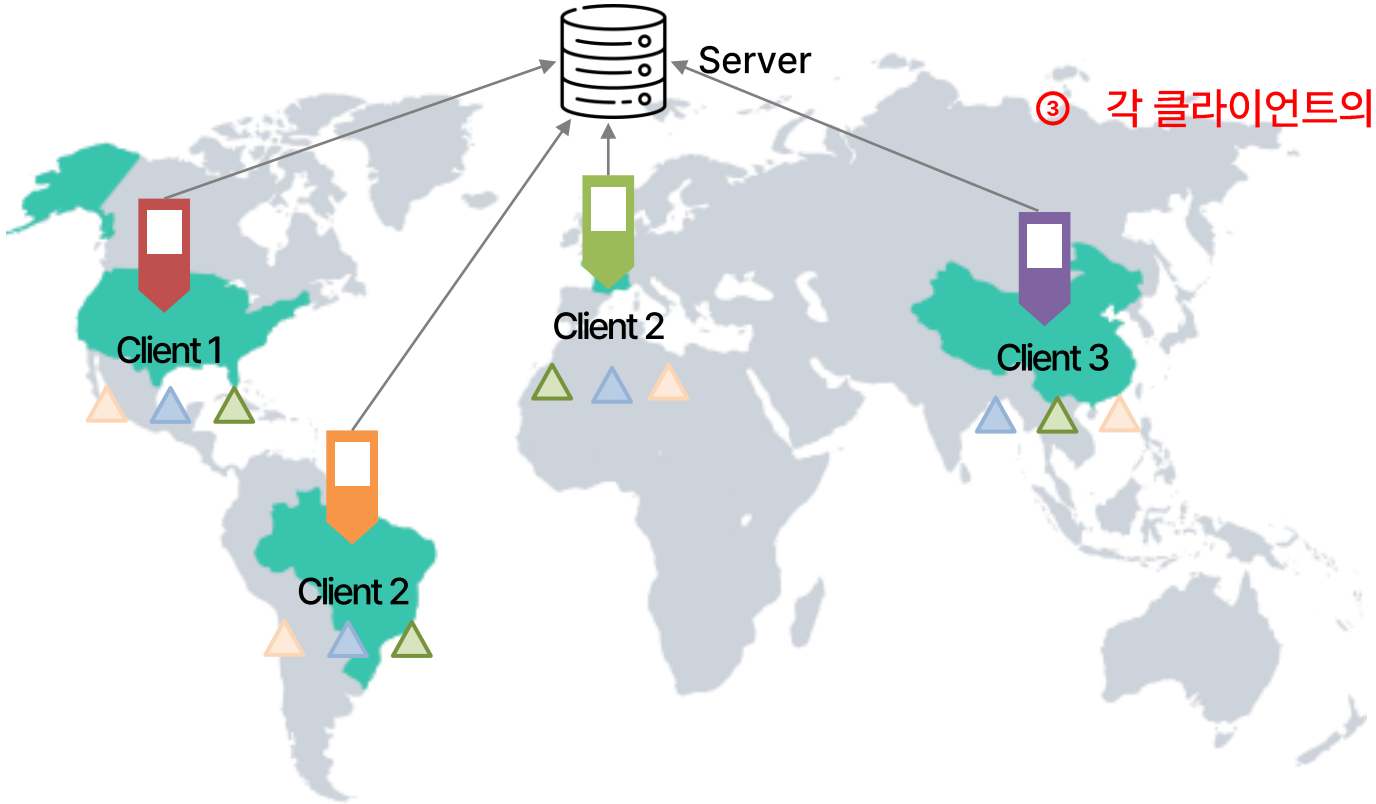


Introduction

Vision Language Models-based Prompt Tuning for Federated Learning

❖ Federated Learning

- 데이터 프라이버시를 보호하기 위해 각 클라이언트들이 자신의 데이터를 공유하지 않고,
  로컬에서 학습한 모델 파라미터만을 중앙 서버와 공유하여 협력적으로 글로벌 모델을 만드는 프레임워크



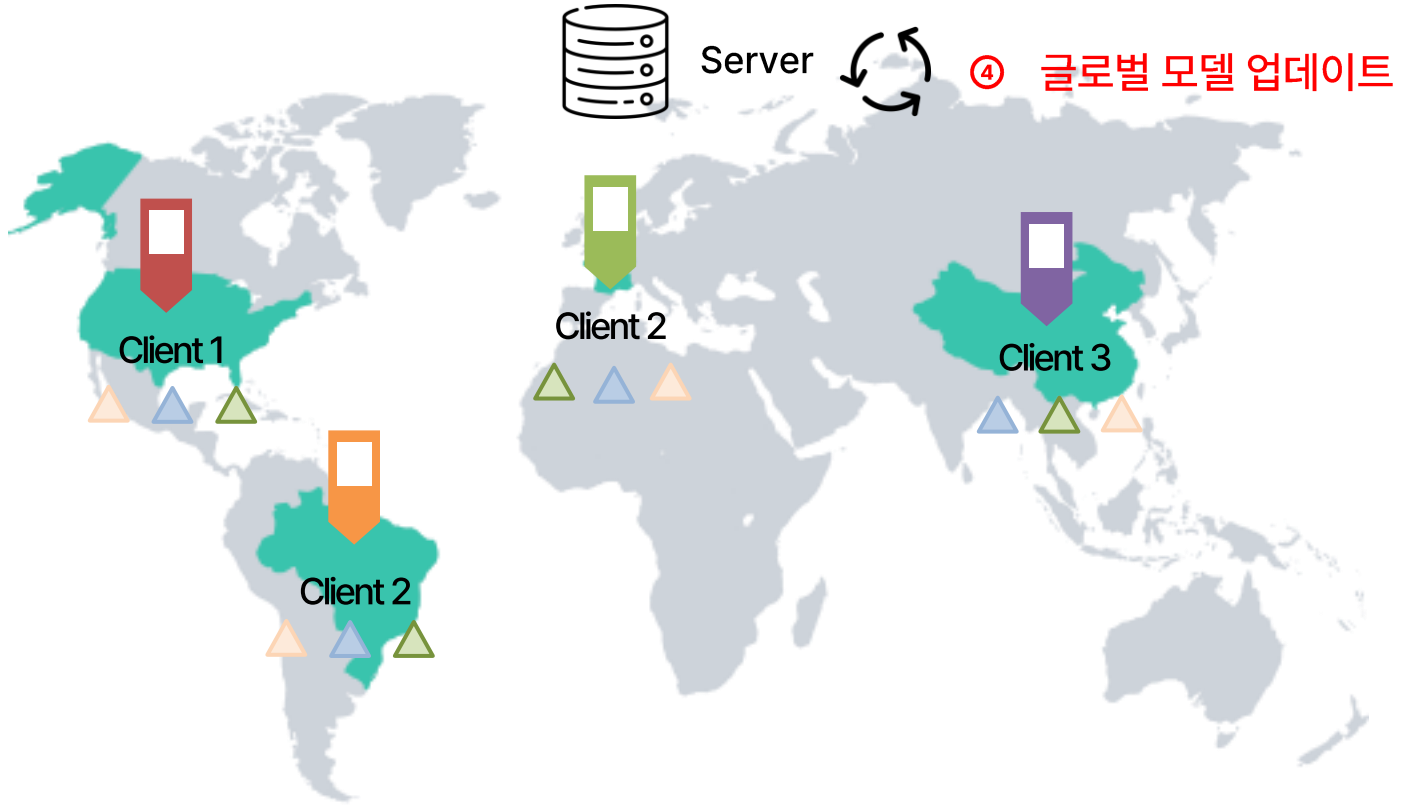
③ 각 클라이언트의 로컬 모델 파라미터 업로드

Introduction

Vision Language Models-based Prompt Tuning for Federated Learning

❖ Federated Learning

- 데이터 프라이버시를 보호하기 위해 각 클라이언트들이 자신의 데이터를 공유하지 않고,
  로컬에서 학습한 모델 파라미터만을 중앙 서버와 공유하여 협력적으로 글로벌 모델을 만드는 프레임워크



Introduction

Vision Language Models-based Prompt Tuning for Federated Learning

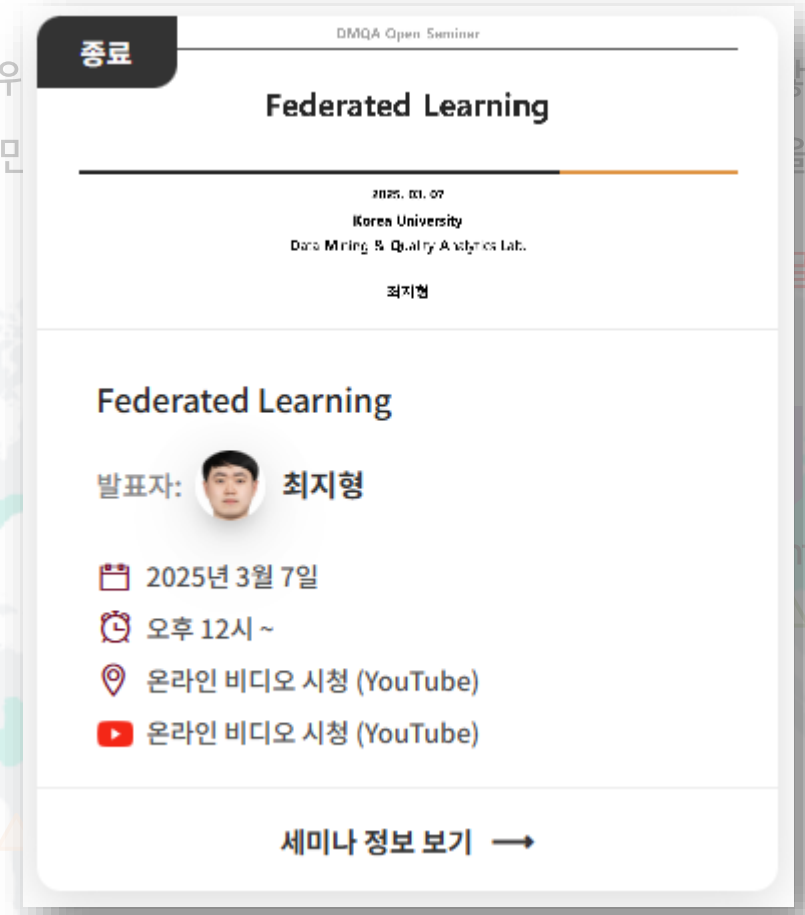
❖ Federated Learning

- 데이터 프라이버시를 보호하기 위
- 로컬에서 학습한 모델 파라미터만

고,

을 만드는 프레임워크

글로벌 모델 업데이트

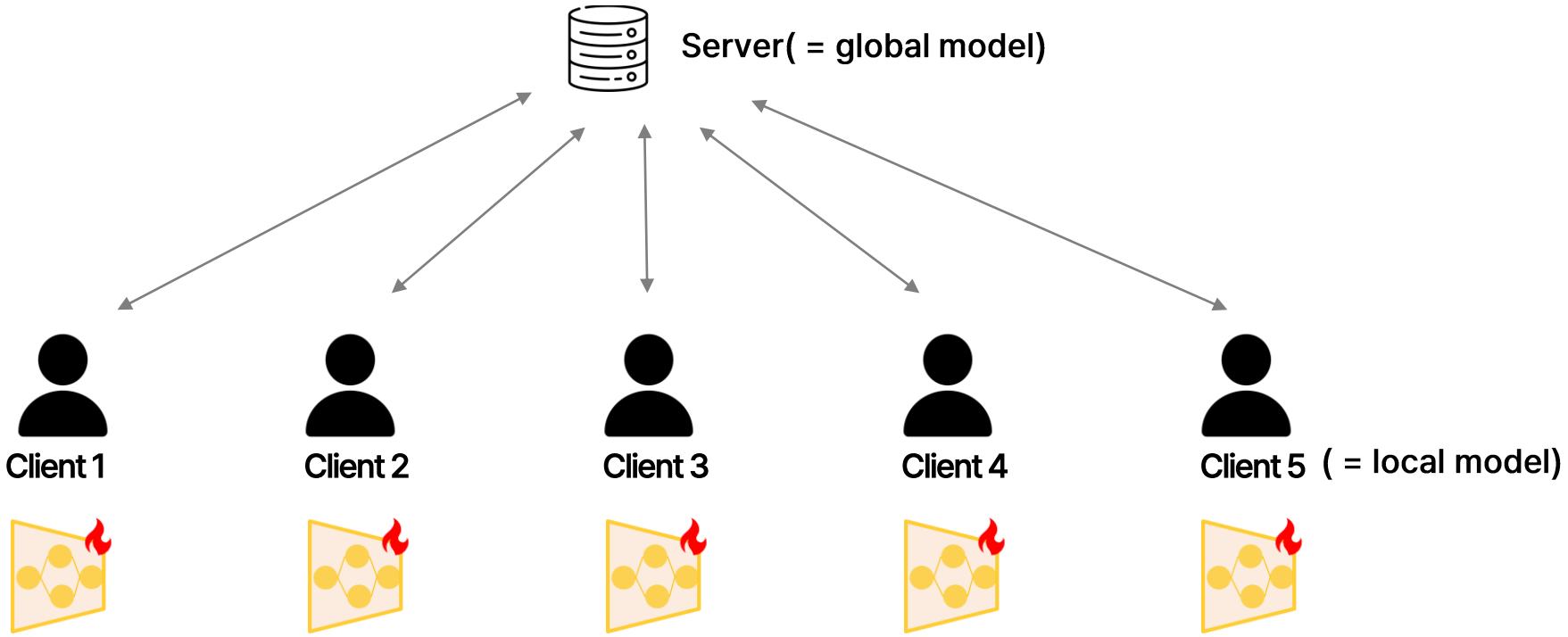


Introduction

Vision Language Models-based Prompt Tuning for Federated Learning

❖ Federated Learning

- 기존 Federated Learning의 한계
  - ✓ 모델 파라미터 공유로 인한 높은 통신 비용으로, 주로 소규모 백본에 국한되어 특징 추출 능력 및 성능이 제한됨

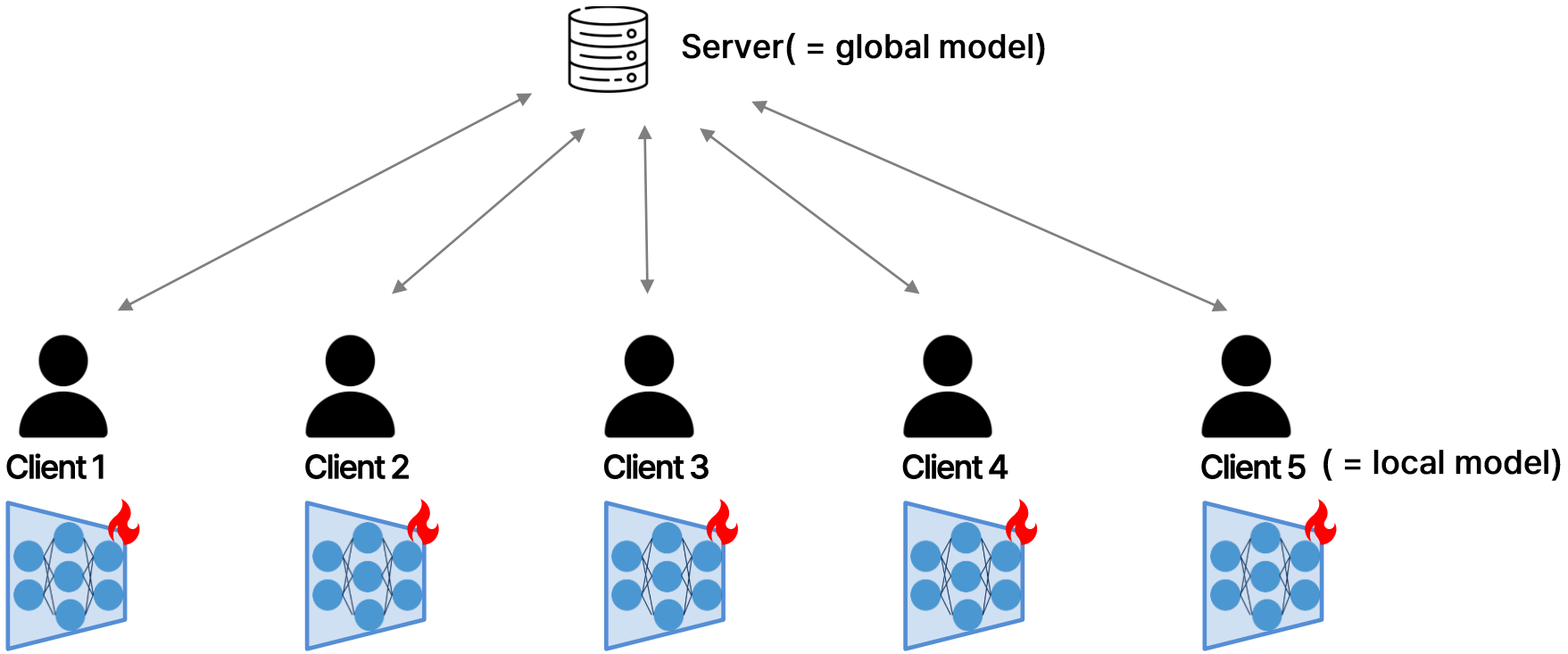


Introduction

Vision Language Models-based Prompt Tuning for Federated Learning

❖ Federated Learning

- Vision-Language Models(VLM)의 잠재력
  - ✓ CLIP과 같은 VLM은 다양한 이미지 분포에서 강건한 특징 표현 학습

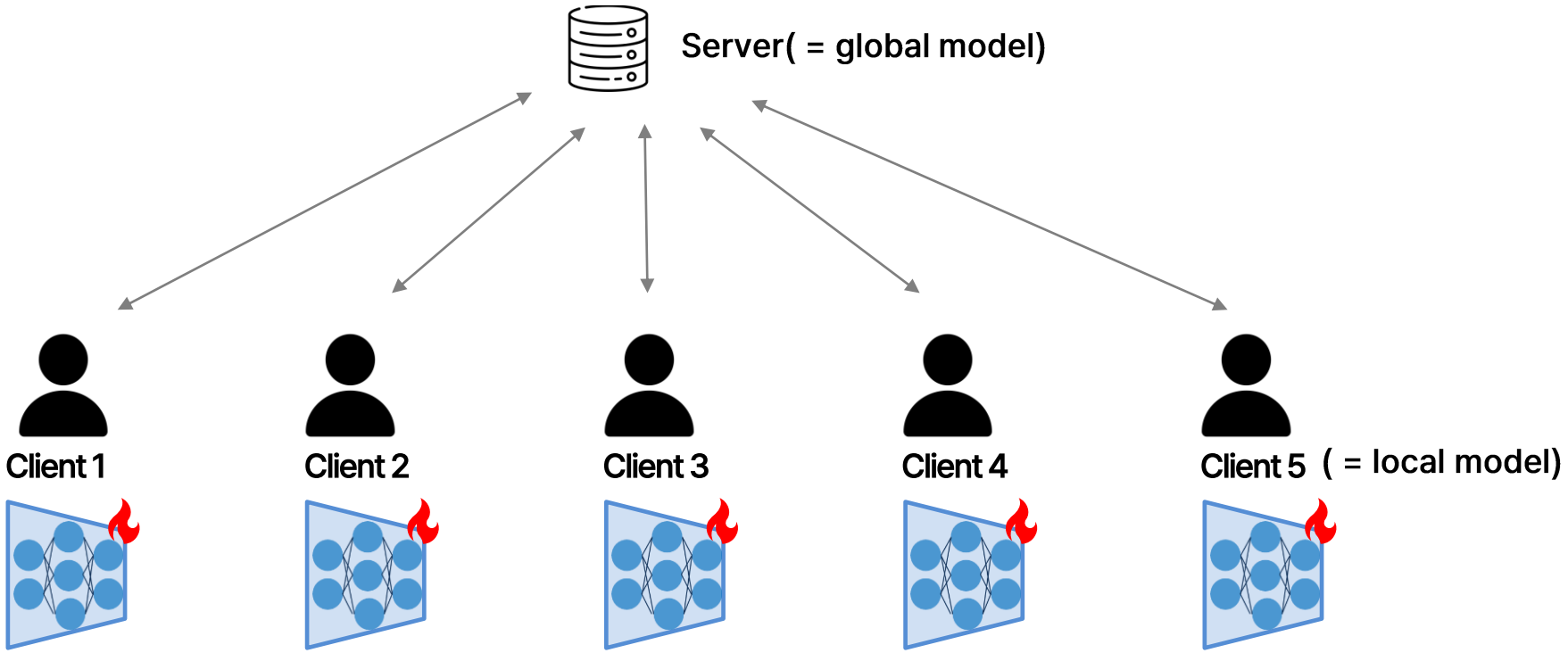


Introduction

Vision Language Models-based Prompt Tuning for Federated Learning

❖ Federated Learning

- Vision-Language Models(VLM)의 잠재력
  - ✓ CLIP과 같은 VLM은 다양한 이미지 분포에서 강건한 특징 표현 학습
  - ✓ But, Federated Learning 환경에 직접 적용 시 → 높은 통신 오버헤드 / 과적합 위험

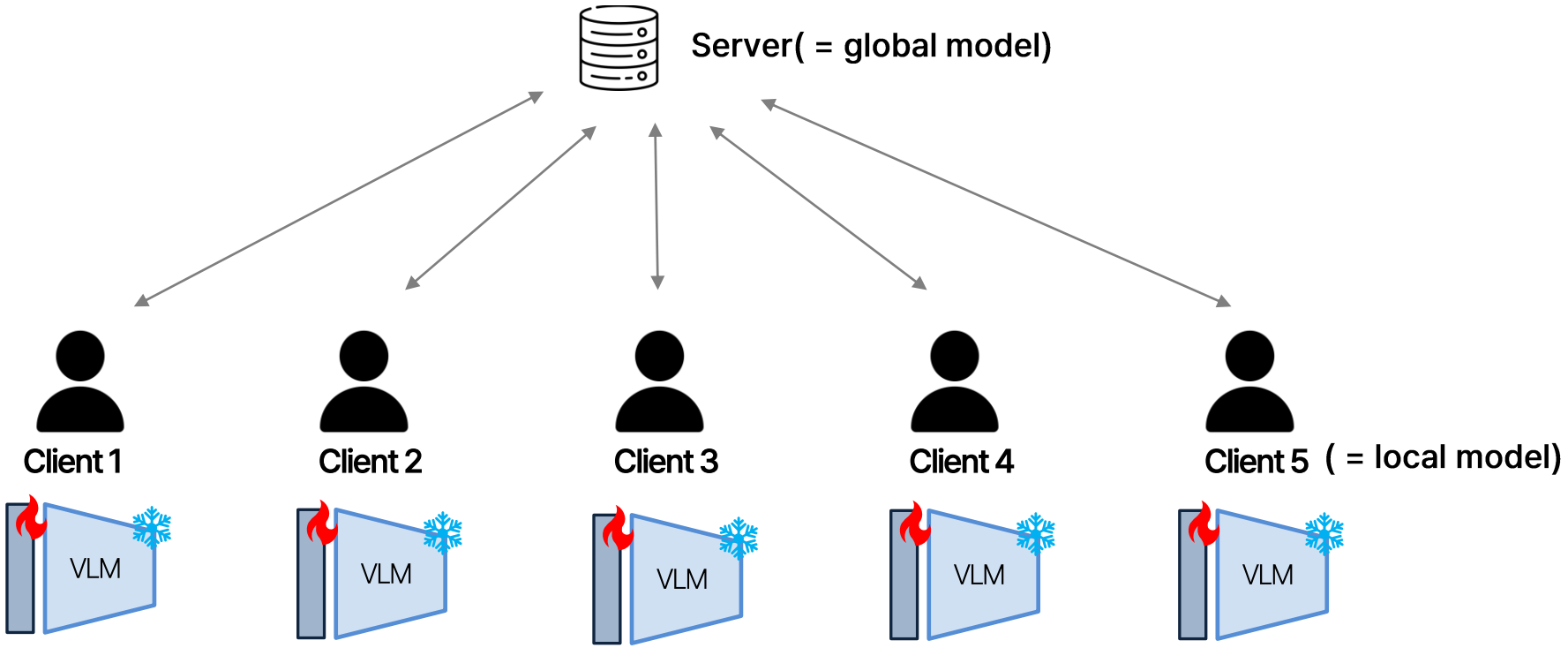


Introduction

Vision Language Models-based Prompt Tuning for Federated Learning

❖ Federated Learning

- 해결 방향: Prompt Learning
  - ✓ 소수의 파라미터만 학습함으로써 통신 비용 절감 및 과적합 방지
  - ✓ VLM의 강건한 사전학습 표현 활용함으로써 데이터 이질성 완화



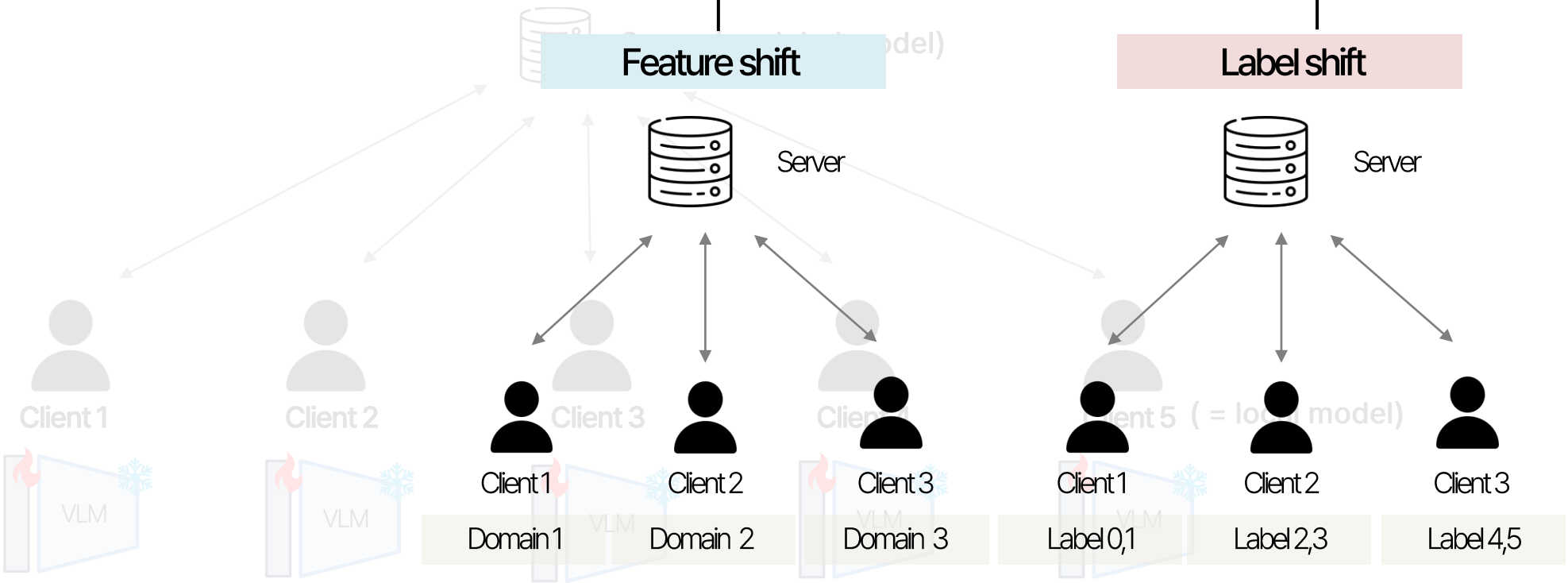
Introduction

Vision Language Models-based Prompt Tuning for Federated Learning

❖ Federated Learning

- 해결 방향: Prompt Learning
  - ✓ 소수의 파라미터만 학습함으로써 통신 비용 절감 및 과적합 방지
  - ✓ VLM의 강건한 사전학습 표현 활용함으로써 데이터 이질성 완화

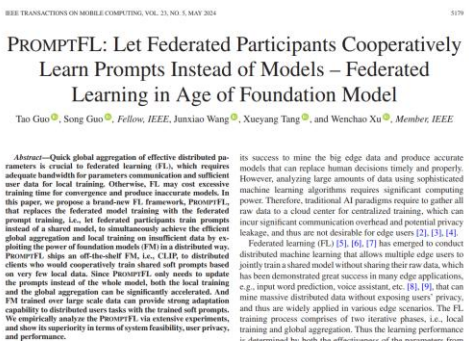
Non-IID로 인해 클라이언트마다 데이터 분포가 달라지는 현상



Related Works

Vision Language Models-based Prompt Tuning for Federated Learning

PromptFL
[IEEE TMC 2024]



IEEE TRANSACTIONS ON MOBILE COMPUTING, VOL. 23, NO. 5, MAY 2024 5179

PROMPTFL: Let Federated Participants Cooperatively Learn Prompts Instead of Models – Federated Learning in Age of Foundation Model

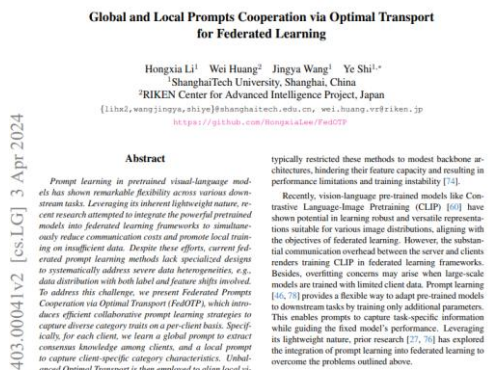
Tao Guo, Song Guo, *Fellow, IEEE*, Junxiao Wang, Xueyang Tang, and Wenchao Xu, *Member, IEEE*

Abstract—Quick global aggregation of effective distributed parameters is crucial to federated learning (FL), which requires adequate bandwidth for parameters communication and sufficient user data for local training. Otherwise, FL may cost excessive training time for convergence and produce inaccurate models. In this paper, we propose a brand-new FL framework, PROMPTFL, that replaces the federated model training with the federated prompt training, i.e., let federated participants train prompts instead of a shared model, to simultaneously achieve the efficient global aggregation and local training on insufficient data by exploiting the power of foundation models (FM) in a distributed way. PROMPTFL ships an off-the-shelf FM, i.e., CLIP, to distributed clients who would cooperatively train shared soft prompts based on very few local data. Since PROMPTFL only needs to update the prompts instead of the whole model, both the local training and the global aggregation can be significantly accelerated. And FM trained over large scale data can provide strong adaptation capability to distributed users tasks with the trained soft prompts. We empirically analyze the PROMPTFL via extensive experiments, and show its superiority in terms of system feasibility, user privacy, and performance.

its success to mine the big edge data and produce accurate models that can replace human decisions timely and properly. However, analyzing large amounts of data using sophisticated machine learning algorithms requires significant computing power. Therefore, traditional AI paradigms require to gather all raw data to a cloud center for centralized training, which can incur significant communication overhead and potential privacy leakage, and thus are not desirable for edge users [2], [3], [4].

Federated learning (FL) [5], [6], [7] has emerged to conduct distributed machine learning that allows multiple edge users to jointly train a shared model without sharing their raw data, which has been demonstrated great success in many edge applications, e.g., input word prediction, voice assistant, etc. [8], [9], that can mine massive distributed data without exposing users' privacy, and thus are widely applied in various edge scenarios. The FL training process comprises of two iterative phases, i.e., local training and global aggregation. Thus the learning performance

FedOTP
[CVPR 2024]



Global and Local Prompts Cooperation via Optimal Transport for Federated Learning

Hongxia Li[1] Wei Huang[2] Jingya Wang[1] Ye Shi[1,*]
[1]ShanghaiTech University, Shanghai, China
[2]RIKEN Center for Advanced Intelligence Project, Japan
{lihx2,wangjingya,shiye}@shanghaitech.edu.cn, wei.huang.vr@riken.jp
https://github.com/HongxiaLee/FedOTP

403.00041v2 [cs.LG] 3 Apr 2024

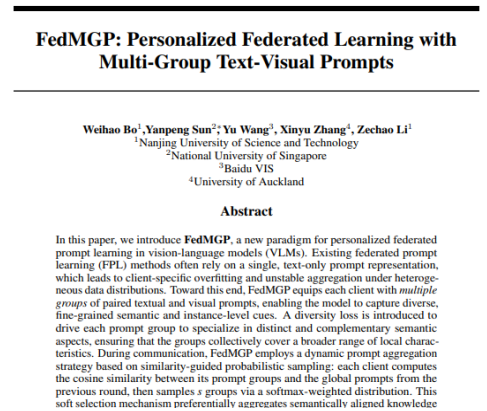
Abstract

Prompt learning in pretrained visual-language models has shown remarkable flexibility across various downstream tasks. Leveraging its inherent lightweight nature, recent research attempted to integrate the powerful pretrained models into federated learning frameworks to simultaneously reduce communication costs and promote local training on insufficient data. Despite these efforts, current federated prompt learning methods lack specialized designs to systematically address severe data heterogeneities, e.g., data distribution with both label and feature shifts involved. To address this challenge, we present Federated Prompts Cooperation via Optimal Transport (FedOTP), which introduces efficient collaborative prompt learning strategies to capture diverse category traits on a per-client basis. Specifically, for each client, we learn a global prompt to extract consensus knowledge among clients, and a local prompt to capture client-specific category characteristics. Unbal-

typically restricted these methods to modest backbone architectures, hindering their feature capacity and resulting in performance limitations and training instability [74].

Recently, vision-language pre-trained models like Contrastive Language-Image Pretraining (CLIP) [60] have shown potential in learning robust and versatile representations suitable for various image distributions, aligning with the objectives of federated learning. However, the substantial communication overhead between the server and clients renders training CLIP in federated learning frameworks. Besides, overfitting concerns may arise when large-scale models are trained with limited client data. Prompt learning [46, 78] provides a flexible way to adapt pre-trained models to downstream tasks by training only additional parameters. This enables prompts to capture task-specific information while guiding the fixed model's performance. Leveraging its lightweight nature, prior research [27, 76] has explored the integration of prompt learning into federated learning to overcome the problems outlined above.

FedMGP
[NeurIPS 2025]



FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

Weihao Bo[1], Yanpeng Sun[2], Yu Wang[3], Xinyu Zhang[4], Zechao Li[1]
[1]Nanjing University of Science and Technology
[2]National University of Singapore
[3]Baidu VIS
[4]University of Auckland

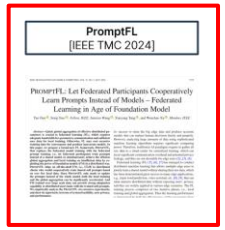
Abstract

In this paper, we introduce **FedMGP**, a new paradigm for personalized federated prompt learning in vision-language models (VLMs). Existing federated prompt learning (FPL) methods often rely on a single, text-only prompt representation, which leads to client-specific overfitting and unstable aggregation under heterogeneous data distributions. Toward this end, FedMGP equips each client with *multiple groups* of paired textual and visual prompts, enabling the model to capture diverse, fine-grained semantic and instance-level cues. A diversity loss is introduced to drive each prompt group to specialize in distinct and complementary semantic aspects, ensuring that the groups collectively cover a broader range of local characteristics. During communication, FedMGP employs a dynamic prompt aggregation strategy based on similarity-guided probabilistic sampling: each client computes the cosine similarity between its prompt groups and the global prompts from the previous round, then samples s groups via a softmax-weighted distribution. This soft selection mechanism preferentially aggregates semantically aligned knowledge

[1] Guo, T., Guo, S., Wang, J., Tang, X., & Xu, W. (2023). PromptFL: Let federated participants cooperatively learn prompts instead of models – federated learning in age of foundation model. IEEE Transactions on Mobile Computing.
[2] Li, H., Diao, E., He, C., Li, Z., Zhao, L., Song, W., Zhu, J., Zhang, S., & Daumé III, H. (2024). Global and local prompts cooperation via optimal transport for federated learning. In CVPR.
[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

PromptFL

Vision Language Models-based Prompt Tuning for Federated Learning



❖ PromptFL: Let Federated Participants Cooperatively Learn Prompts Instead of Models

- 2024년 IEEE Transactions on Mobile Computing, 인용 수 270회
- 기존 FL은 전체 모델 파라미터를 공유하여 높은 통신 비용 발생, VLM을 FL에 직접 적용하기 어려움
- 텍스트 프롬프트만을 클라이언트 간 학습·집계함으로써 소수의 파라미터만으로 FL 수행

PROMPTFL: Let Federated Participants Cooperatively Learn Prompts Instead of Models — Federated Learning in Age of Foundation Model

Tao Guo, Song Guo, Junxiao Wang, Wenchao Xu

Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

Abstract

Quick global aggregation of effective distributed parameters is crucial to federated learning (FL), which requires adequate bandwidth for parameters communication and sufficient user data for local training. Otherwise, FL may cost excessive training time for convergence and produce inaccurate models. In this paper, we propose a brand-new FL framework, PROMPTFL, that replaces the federated model training with the federated prompt training, i.e., let federated participants train prompts instead of a shared model, to simultaneously achieve the efficient global aggregation and local training on insufficient data by exploiting the power of foundation models (FM) in a distributed way. PROMPTFL ships an off-the-shelf FM, i.e., CLIP, to distributed clients who would cooperatively train shared soft prompts based on very few local data. Since PROMPTFL only needs to update the prompts instead of the whole model, both the local training and the global aggregation from local training and smooth aggregation of them. However, these two requirements are not easy to satisfy in edge environment, i.e., edge users often have limited bandwidth and insufficient data, which can cause inefficient parameters aggregation, excessive training time and reduced model accuracy.

Existing research efforts have focused on improving the FL optimization process (Li et al. 2020; Zhao et al. 2018) or refining model architectures (Qu et al. 2022), but this does not change that FL inherently entails a large number of communication rounds and a large amount of labeled data for training, which are often unavailable for edge users. Such challenges are particularly salient under the combined effect of a long training process and unfavorable factors such as non-IID and unbalanced data, limited communication bandwidth, and unreliable and limited device availability.

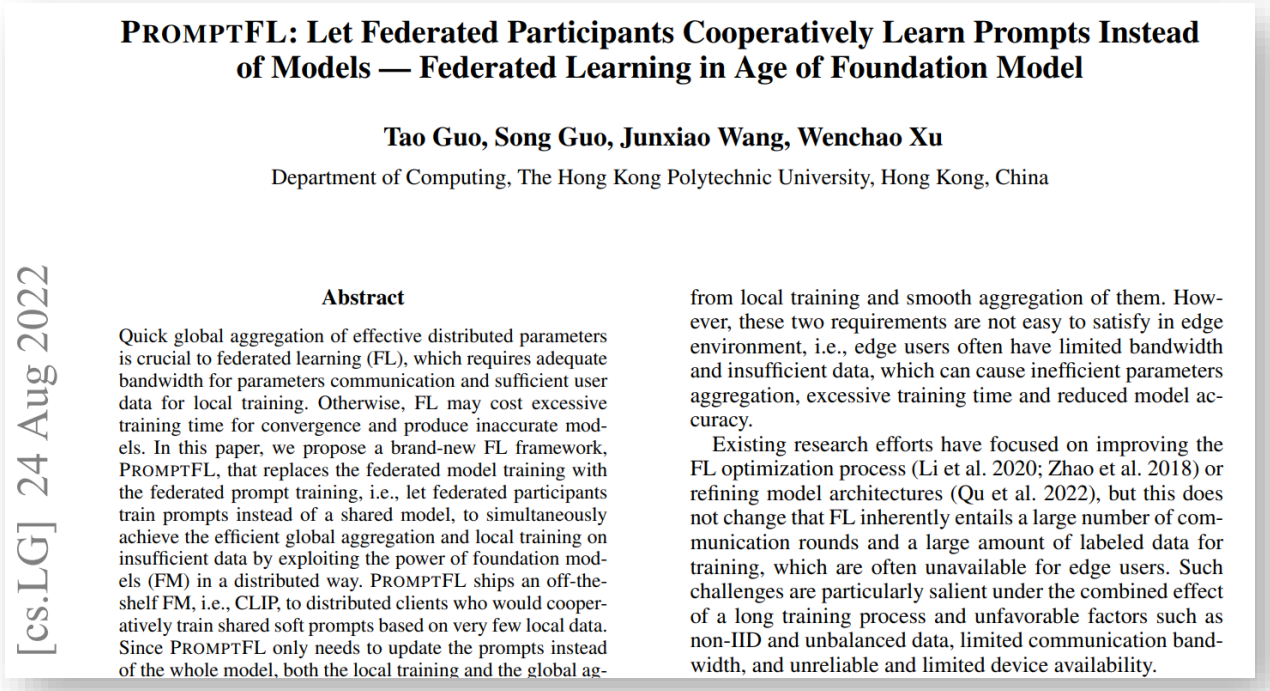
[1] Guo, T., Guo, S., Wang, J., Tang, X., & Xu, W. (2023). PromptFL: Let federated participants cooperatively learn prompts instead of models – federated learning in age of foundation model. IEEE Transactions on Mobile Computing.

PromptFL

Vision Language Models-based Prompt Tuning for Federated Learning

❖ PromptFL: Let Federated Participants Cooperatively Learn Prompts Instead of Models

- 2024년 IEEE Transactions on Mobile Computing, 인용 수 270회
- 기존 FL은 전체 모델 파라미터를 공유하여 높은 통신 비용 발생, VLM을 FL에 직접 적용하기 어려움
- 텍스트 프롬프트만을 클라이언트 간 학습·집계함으로써 소수의 파라미터만으로 FL 수행



PROMPTFL: Let Federated Participants Cooperatively Learn Prompts Instead of Models — Federated Learning in Age of Foundation Model

Tao Guo, Song Guo, Junxiao Wang, Wenchao Xu

Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

[cs.LG] 24 Aug 2022

Abstract

Quick global aggregation of effective distributed parameters is crucial to federated learning (FL), which requires adequate bandwidth for parameters communication and sufficient user data for local training. Otherwise, FL may cost excessive training time for convergence and produce inaccurate models. In this paper, we propose a brand-new FL framework, PROMPTFL, that replaces the federated model training with the federated prompt training, i.e., let federated participants train prompts instead of a shared model, to simultaneously achieve the efficient global aggregation and local training on insufficient data by exploiting the power of foundation models (FM) in a distributed way. PROMPTFL ships an off-the-shelf FM, i.e., CLIP, to distributed clients who would cooperatively train shared soft prompts based on very few local data. Since PROMPTFL only needs to update the prompts instead of the whole model, both the local training and the global aggregation from local training and smooth aggregation of them. However, these two requirements are not easy to satisfy in edge environment, i.e., edge users often have limited bandwidth and insufficient data, which can cause inefficient parameters aggregation, excessive training time and reduced model accuracy.

Existing research efforts have focused on improving the FL optimization process (Li et al. 2020; Zhao et al. 2018) or refining model architectures (Qu et al. 2022), but this does not change that FL inherently entails a large number of communication rounds and a large amount of labeled data for training, which are often unavailable for edge users. Such challenges are particularly salient under the combined effect of a long training process and unfavorable factors such as non-IID and unbalanced data, limited communication bandwidth, and unreliable and limited device availability.

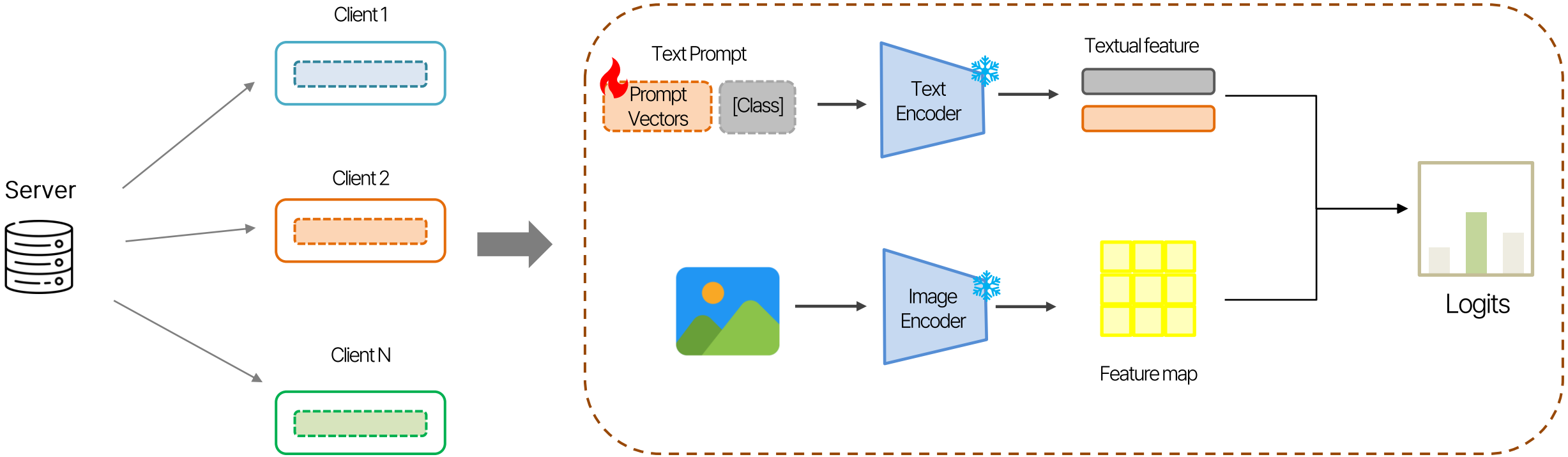
[1] Guo, T., Guo, S., Wang, J., Tang, X., & Xu, W. (2023). PromptFL: Let federated participants cooperatively learn prompts instead of models – federated learning in age of foundation model. IEEE Transactions on Mobile Computing.

PromptFL

Vision Language Models-based Prompt Tuning for Federated Learning

❖ PromptFL: Let Federated Participants Cooperatively Learn Prompts Instead of Models

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 텍스트 프롬프트 벡터만을 로컬에서 학습
- 각 클라이언트의 텍스트 프롬프트 벡터를 서버에서 가중 평균(FedAVG)으로 집계



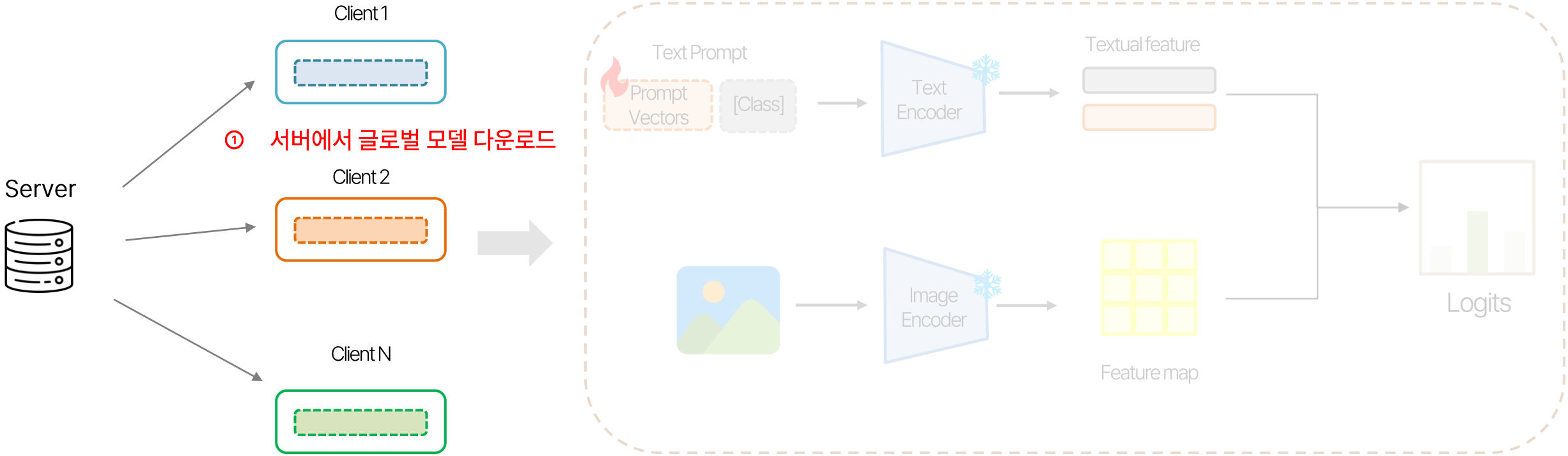
[1] Guo, T., Guo, S., Wang, J., Tang, X., & Xu, W. (2023). PromptFL: Let federated participants cooperatively learn prompts instead of models – federated learning in age of foundation model. IEEE Transactions on Mobile Computing.

PromptFL

Vision Language Models-based Prompt Tuning for Federated Learning

❖ PromptFL: Let Federated Participants Cooperatively Learn Prompts Instead of Models

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 텍스트 프롬프트 벡터만을 로컬에서 학습
- 각 클라이언트의 텍스트 프롬프트 벡터를 서버에서 가중 평균(FedAVG)으로 집계



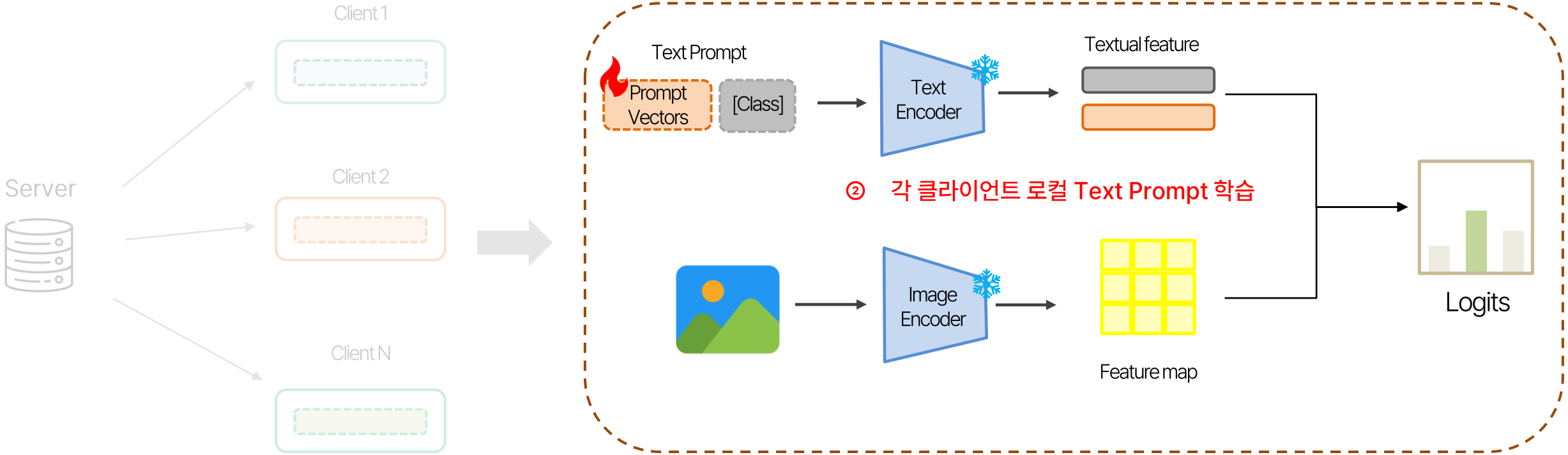
[1] Guo, T., Guo, S., Wang, J., Tang, X., & Xu, W. (2023). PromptFL: Let federated participants cooperatively learn prompts instead of models – federated learning in age of foundation model. IEEE Transactions on Mobile Computing.

PromptFL

Vision Language Models-based Prompt Tuning for Federated Learning

❖ PromptFL: Let Federated Participants Cooperatively Learn Prompts Instead of Models

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 텍스트 프롬프트 벡터만을 로컬에서 학습
- 각 클라이언트의 텍스트 프롬프트 벡터를 서버에서 가중 평균(FedAVG)으로 집계



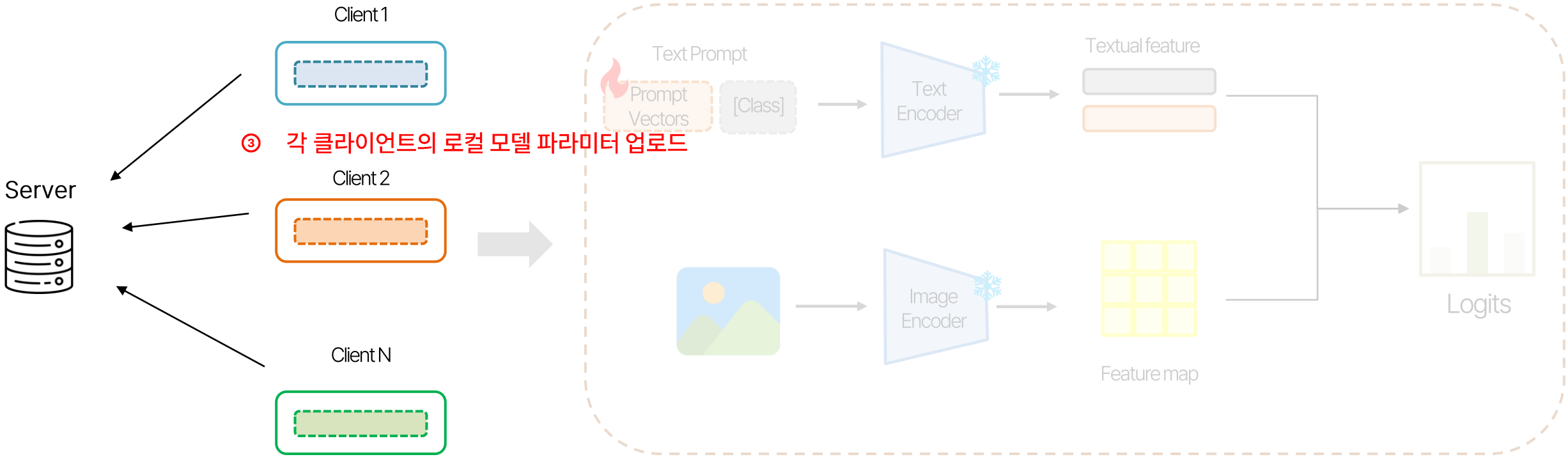
[1] Guo, T., Guo, S., Wang, J., Tang, X., & Xu, W. (2023). PromptFL: Let federated participants cooperatively learn prompts instead of models – federated learning in age of foundation model. IEEE Transactions on Mobile Computing.

PromptFL

Vision Language Models-based Prompt Tuning for Federated Learning

❖ PromptFL: Let Federated Participants Cooperatively Learn Prompts Instead of Models

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 텍스트 프롬프트 벡터만을 로컬에서 학습
- 각 클라이언트의 텍스트 프롬프트 벡터를 서버에서 가중 평균(FedAVG)으로 집계



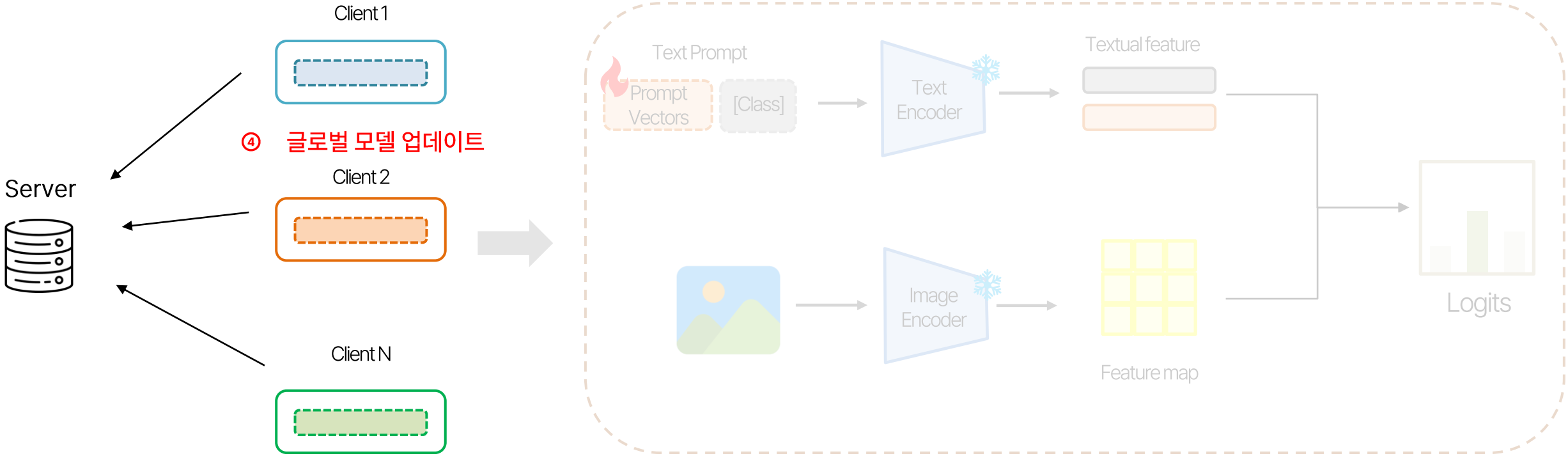
[1] Guo, T., Guo, S., Wang, J., Tang, X., & Xu, W. (2023). PromptFL: Let federated participants cooperatively learn prompts instead of models – federated learning in age of foundation model. IEEE Transactions on Mobile Computing.

PromptFL

Vision Language Models-based Prompt Tuning for Federated Learning

❖ PromptFL: Let Federated Participants Cooperatively Learn Prompts Instead of Models

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 텍스트 프롬프트 벡터만을 로컬에서 학습
- 각 클라이언트의 텍스트 프롬프트 벡터를 서버에서 가중 평균(FedAVG)으로 집계



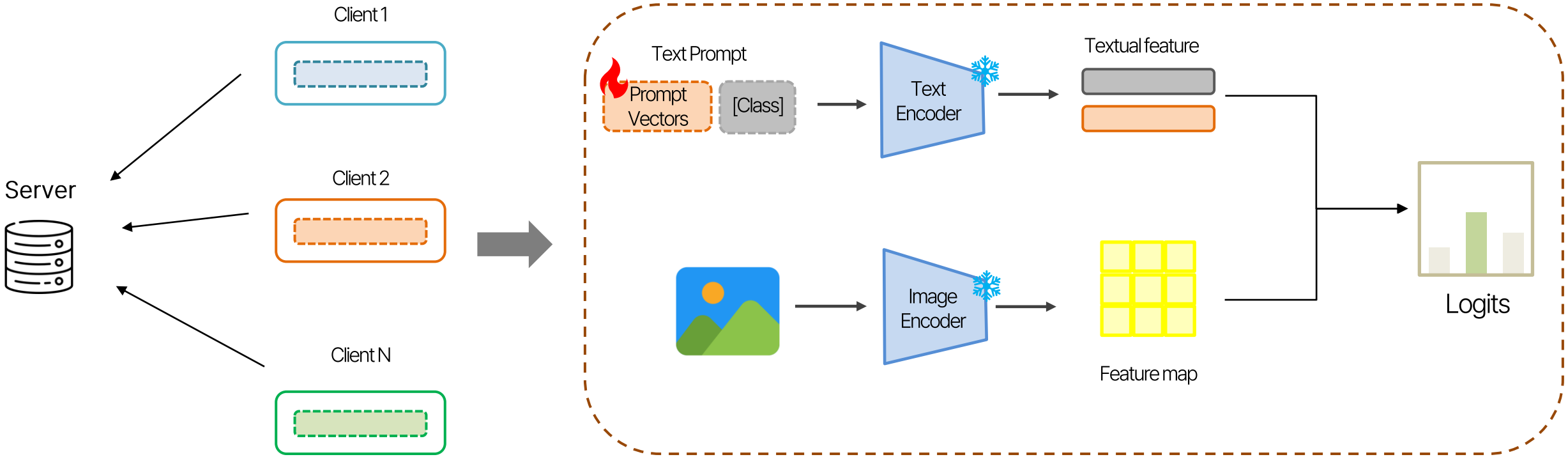
[1] Guo, T., Guo, S., Wang, J., Tang, X., & Xu, W. (2023). PromptFL: Let federated participants cooperatively learn prompts instead of models – federated learning in age of foundation model. IEEE Transactions on Mobile Computing.

PromptFL

Vision Language Models-based Prompt Tuning for Federated Learning

❖ PromptFL: Let Federated Participants Cooperatively Learn Prompts Instead of Models

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 텍스트 프롬프트 벡터만을 로컬에서 학습
- 각 클라이언트의 텍스트 프롬프트 벡터를 서버에서 가중 평균(FedAVG)으로 집계



[1] Guo, T., Guo, S., Wang, J., Tang, X., & Xu, W. (2023). PromptFL: Let federated participants cooperatively learn prompts instead of models – federated learning in age of foundation model. IEEE Transactions on Mobile Computing.

PromptFL

Vision Language Models-based Prompt Tuning for Federated Learning

❖ Experiments

- IID / Non-IID 환경에서 성능 비교
- PromptFL은 전체 파라미터의 0.01%만 학습함에도 Non-IID에서 Finetuning FL 대비 압도적 성능 유지

Benchmark	Method	IID				Extreme Non-IID				Learnable Parameters	
		Accuracy↑		*F-1*↑		*Accuracy*↑		*F-1*↑			
		Rn50	Vit	Rn50	Vit	Rn50	Vit	Rn50	Vit	Rn50	Vit
Caltech101	PromptFL	**90.18**	**94.65**	**86.09**	**91.76**	**88.72**	**94.12**	**83.98**	**90.48**	0.1%	0.01%
	Finetuning FL	90.02	93.1	84.72	89.07	29.78	29.89	12.2	12.2	100%	100%
	FL from scratch	32.41	32.49	10.51	12.89	-	-	-	-	100%	100%
Flowers102	PromptFL	88.14	90.5	87.62	90.14	**66.3**	**74.75**	**60.14**	**69.13**	0.1%	0.01%
	Finetuning FL	**92.6**	**91.9**	**91.56**	**90.7**	24.4	24.5	10.68	11.18	100%	100%
	FL from scratch	33.17	38	25.7	32.5	-	-	-	-	100%	100%
OxfordPets	PromptFL	88.5	**92.89**	88.44	**92.8**	**87.03**	**89.51**	**86.85**	**88.45**	0.1%	0.01%
	Finetuning FL	**90.38**	92.1	**90.06**	91.92	24.83	25.27	11.3	11.93	100%	100%
	FL from scratch	10.25	8.722	7.624	8.318	-	-	-	-	100%	100%
Food101	PromptFL	**78.0**	**85.75**	**77.9**	**85.66**	**78.1**	**85.88**	**78.03**	**85.8**	0.1%	0.01%
	Finetuning FL	69.28	76.68	69.08	76.85	22.92	23.8	10.19	10.73	100%	100%
	FL from scratch	21.11	21.03	19.75	19.92	-	-	-	-	100%	100%

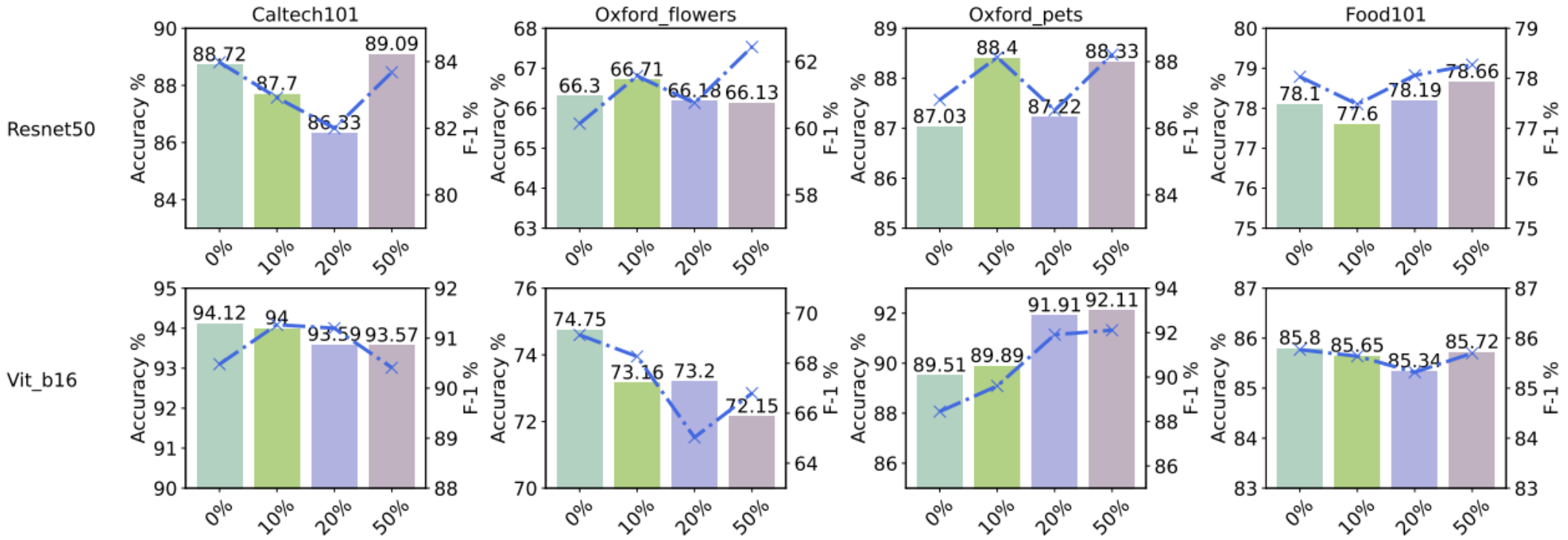
[1] Guo, T., Guo, S., Wang, J., Tang, X., & Xu, W. (2023). PromptFL: Let federated participants cooperatively learn prompts instead of models – federated learning in age of foundation model. IEEE Transactions on Mobile Computing.

PromptFL

Vision Language Models-based Prompt Tuning for Federated Learning

❖ Experiments

- 클라이언트 간 클래스 중복 비율(0% → 50%)에 따른 성능 변화 분석
- 클래스 분포가 극단적으로 달라져도 성능이 안정적으로 유지됨 → Label Shift에 강건함



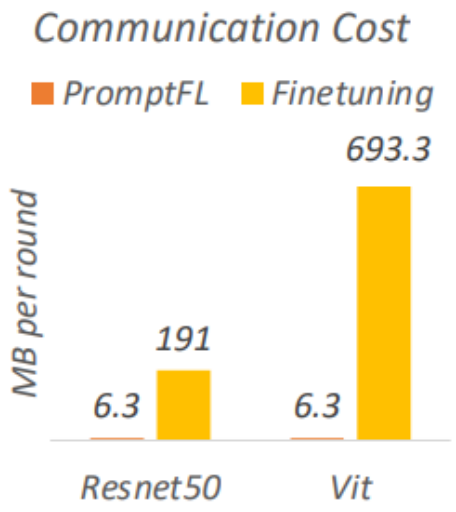
[1] Guo, T., Guo, S., Wang, J., Tang, X., & Xu, W. (2023). PromptFL: Let federated participants cooperatively learn prompts instead of models – federated learning in age of foundation model. IEEE Transactions on Mobile Computing.

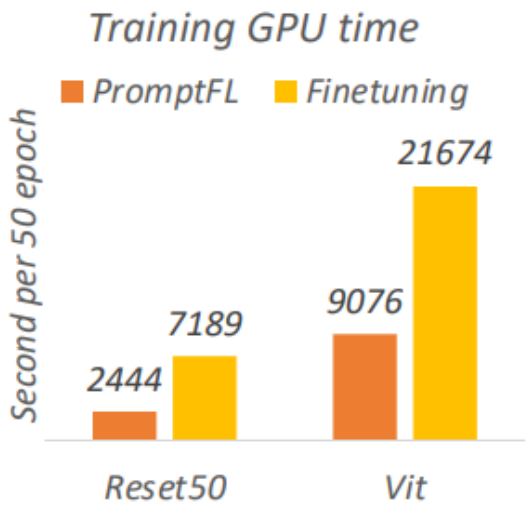
PromptFL

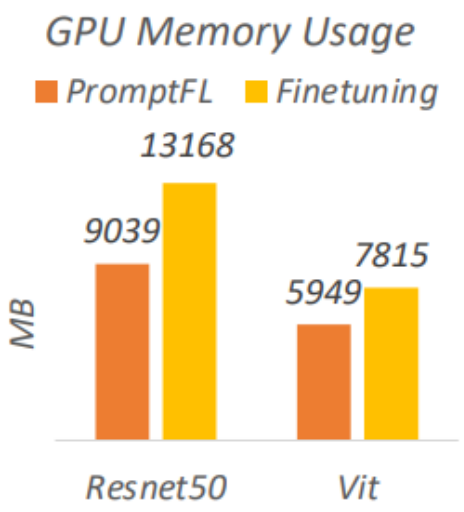
Vision Language Models-based Prompt Tuning for Federated Learning

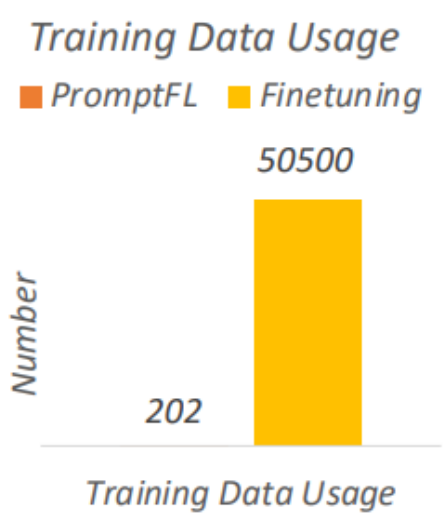
❖ Experiments

- PromptFL vs Finetuning FL 간 통신 비용, GPU 학습 시간, 메모리 사용량, 학습 데이터 사용량 비교









[1] Guo, T., Guo, S., Wang, J., Tang, X., & Xu, W. (2023). PromptFL: Let federated participants cooperatively learn prompts instead of models – federated learning in age of foundation model. IEEE Transactions on Mobile Computing.

FedOTP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedOTP: Global and local prompts cooperation via optimal transport for federated learning

- 2024년 CVPR, 인용 수 90회
- FL 환경에서 클라이언트 간 심각한 데이터 이질성 문제를 체계적으로 해결하는 설계 부재
- Global/Local 이중 텍스트 프롬프트 구조 + Unbalanced OT 정렬로 공통 지식과 클라이언트별 개인화를 동시에 달성

Global and Local Prompts Cooperation via Optimal Transport for Federated Learning

Hongxia Li[1] Wei Huang[2] Jingya Wang[1] Ye Shi[1,*]
[1]ShanghaiTech University, Shanghai, China
[2]RIKEN Center for Advanced Intelligence Project, Japan
{lihx2,wangjingya,shiye}@shanghaitech.edu.cn, wei.huang.vr@riken.jp
https://github.com/HongxiaLee/FedOTP

Abstract

Prompt learning in pretrained visual-language models has shown remarkable flexibility across various downstream tasks. Leveraging its inherent lightweight nature, recent research attempted to integrate the powerful pretrained models into federated learning frameworks to simultaneously reduce communication costs and promote local training on insufficient data. Despite these efforts, current federated prompt learning methods lack specialized designs to systematically address severe data heterogeneities, e.g.,

typically restricted these methods to modest backbone architectures, hindering their feature capacity and resulting in performance limitations and training instability [74].

Recently, vision-language pre-trained models like Contrastive Language-Image Pretraining (CLIP) [60] have shown potential in learning robust and versatile representations suitable for various image distributions, aligning with the objectives of federated learning. However, the substantial communication overhead between the server and clients renders training CLIP in federated learning frameworks. Besides, overfitting concerns may arise when large-scale

[2] Li, H., Diao, E., He, C., Li, Z., Zhao, L., Song, W., Zhu, J., Zhang, S., & Daumé III, H. (2024). Global and local prompts cooperation via optimal transport for federated learning. In CVPR.

FedOTP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedOTP: Global and local prompts cooperation via optimal transport for federated learning

- 2024년 CVPR, 인용 수 90회
- FL 환경에서 클라이언트 간 심각한 데이터 이질성 문제를 체계적으로 해결하는 설계 부재
- Global/Local 이중 텍스트 프롬프트 구조 + Unbalanced OT 정렬로 공통 지식과 클라이언트별 개인화를 동시에 달성

Global and Local Prompts Cooperation via Optimal Transport for Federated Learning

Hongxia Li[1] Wei Huang[2] Jingya Wang[1] Ye Shi[1,*]
[1]ShanghaiTech University, Shanghai, China
[2]RIKEN Center for Advanced Intelligence Project, Japan
{lihx2,wangjingya,shiye}@shanghaitech.edu.cn, wei.huang.vr@riken.jp
https://github.com/HongxiaLee/FedOTP

Abstract

Prompt learning in pretrained visual-language models has shown remarkable flexibility across various downstream tasks. Leveraging its inherent lightweight nature, recent research attempted to integrate the powerful pretrained models into federated learning frameworks to simultaneously reduce communication costs and promote local training on insufficient data. Despite these efforts, current federated prompt learning methods lack specialized designs to systematically address severe data heterogeneities, e.g.,

typically restricted these methods to modest backbone architectures, hindering their feature capacity and resulting in performance limitations and training instability [74].

Recently, vision-language pre-trained models like Contrastive Language-Image Pretraining (CLIP) [60] have shown potential in learning robust and versatile representations suitable for various image distributions, aligning with the objectives of federated learning. However, the substantial communication overhead between the server and clients renders training CLIP in federated learning frameworks. Besides, overfitting concerns may arise when large-scale

[2] Li, H., Diao, E., He, C., Li, Z., Zhao, L., Song, W., Zhu, J., Zhang, S., & Daumé III, H. (2024). Global and local prompts cooperation via optimal transport for federated learning. In CVPR.

FedOTP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedOTP: Global and local prompts cooperation via optimal transport for federated learning

- 2024년 CVPR, 인용 수 90회
- FL 환경에서 클라이언트 간 심각한 데이터 이질성 문제를 체계적으로 해결하는 설계 부재
- Global/Local 이중 텍스트 프롬프트 구조 + Unbalanced OT 정렬로 공통 지식과 클라이언트별 개인화를 동시에 달성

Global and Local Prompts Cooperation via Optimal Transport for Federated Learning

Hongxia Li[1] Wei Huang[2] Jingya Wang[1] Ye Shi[1,*]
[1]ShanghaiTech University, Shanghai, China
[2]RIKEN Center for Advanced Intelligence Project, Japan
{lihx2,wangjingya,shiye}@shanghaitech.edu.cn, wei.huang.vr@riken.jp
https://github.com/HongxiaLee/FedOTP

Abstract

Prompt learning in pretrained visual-language models has shown remarkable flexibility across various downstream tasks. Leveraging its inherent lightweight nature, recent research attempted to integrate the powerful pretrained models into federated learning frameworks to simultaneously reduce communication costs and promote local training on insufficient data. Despite these efforts, current federated prompt learning methods lack specialized designs to systematically address severe data heterogeneities, e.g.,

typically restricted these methods to modest backbone architectures, hindering their feature capacity and resulting in performance limitations and training instability [74].

Recently, vision-language pre-trained models like Contrastive Language-Image Pretraining (CLIP) [60] have shown potential in learning robust and versatile representations suitable for various image distributions, aligning with the objectives of federated learning. However, the substantial communication overhead between the server and clients renders training CLIP in federated learning frameworks. Besides, overfitting concerns may arise when large-scale

[2] Li, H., Diao, E., He, C., Li, Z., Zhao, L., Song, W., Zhu, J., Zhang, S., & Daumé III, H. (2024). Global and local prompts cooperation via optimal transport for federated learning. In CVPR.

FedOTP

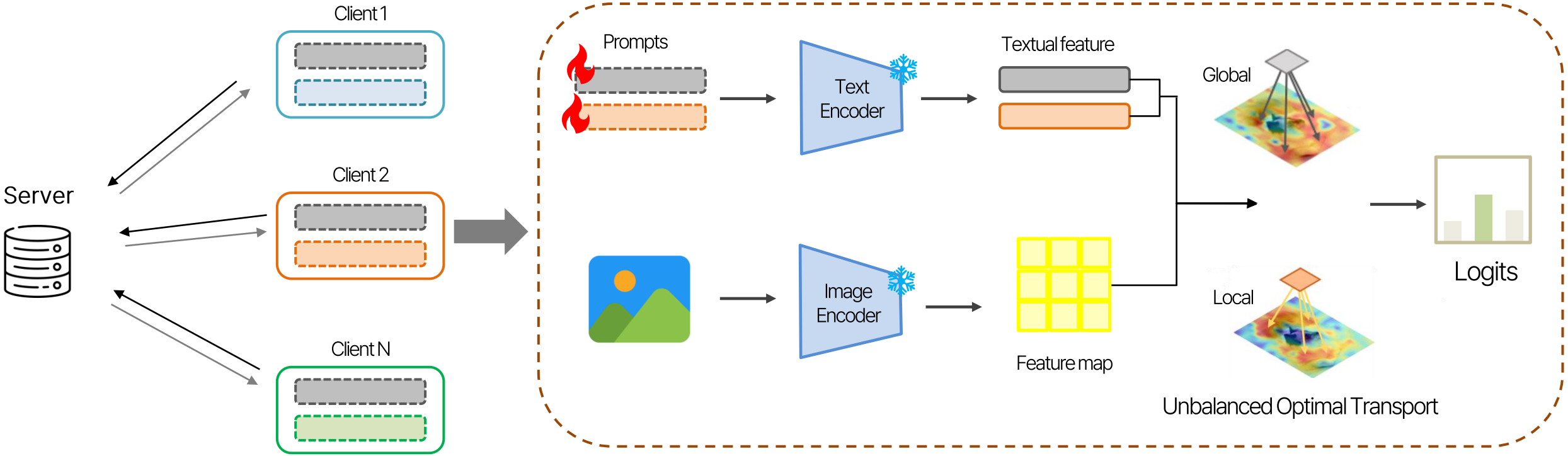
Vision Language Models-based Prompt Tuning for Federated Learning

: Global text prompt

: Local text prompt

❖ FedOTP: Global and local prompts cooperation via optimal transport for federated learning

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Global/Local 이중 텍스트 프롬프트 부여
- Unbalanced Optimal Transport으로 로컬 Visual feature와 Global/Local 프롬프트를 정렬하여 데이터 이질성 극복



[2] Li, H., Diao, E., He, C., Li, Z., Zhao, L., Song, W., Zhu, J., Zhang, S., & Daumé III, H. (2024). Global and local prompts cooperation via optimal transport for federated learning. In CVPR.

FedOTP

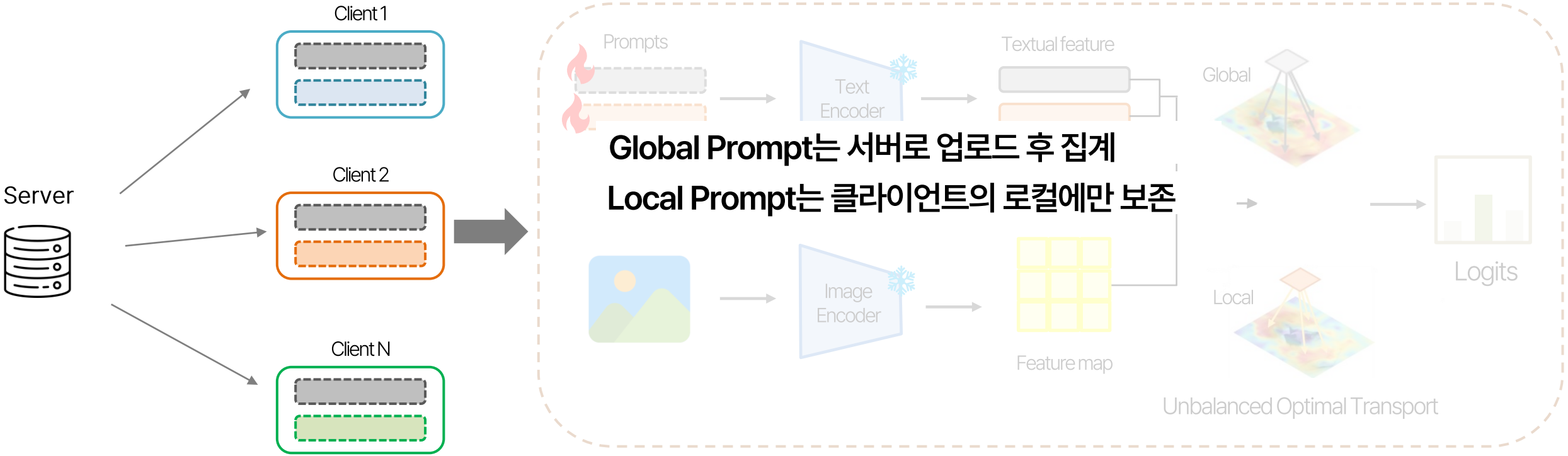
Vision Language Models-based Prompt Tuning for Federated Learning

: Global text prompt

: Local text prompt

❖ FedOTP: Global and local prompts cooperation via optimal transport for federated learning

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Global/Local 이중 텍스트 프롬프트 부여
- Unbalanced Optimal Transport으로 로컬 Visual feature와 Global/Local 프롬프트를 정렬하여 데이터 이질성 극복



[2] Li, H., Diao, E., He, C., Li, Z., Zhao, L., Song, W., Zhu, J., Zhang, S., & Daumé III, H. (2024). Global and local prompts cooperation via optimal transport for federated learning. In CVPR.

FedOTP

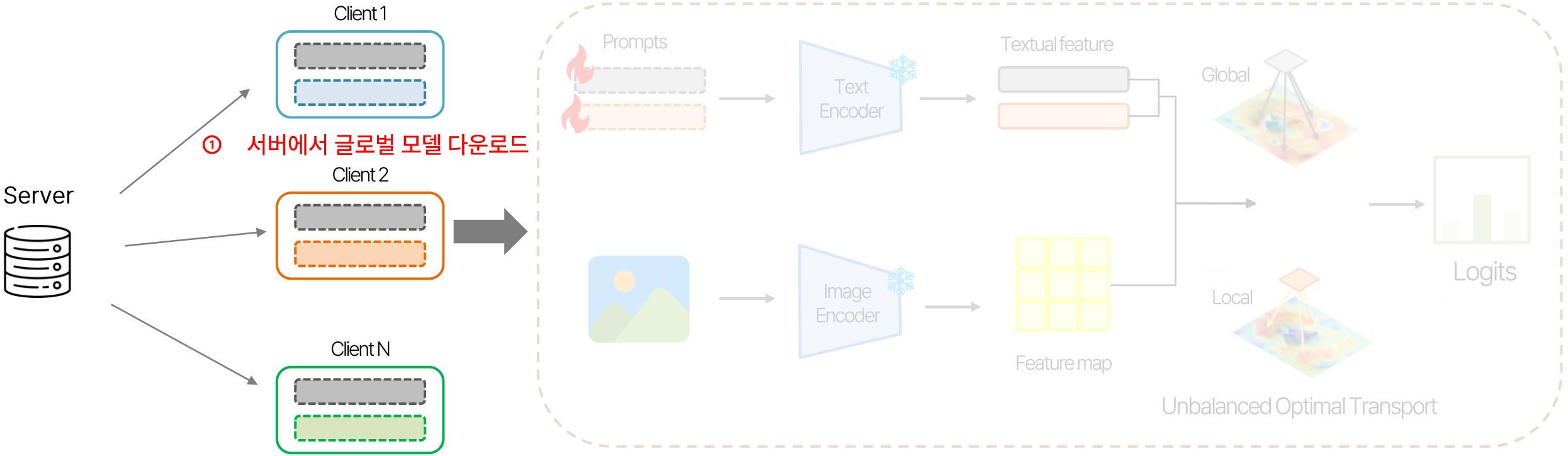
Vision Language Models-based Prompt Tuning for Federated Learning

: Global text prompt

: Local text prompt

❖ FedOTP: Global and local prompts cooperation via optimal transport for federated learning

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Global/Local 이중 텍스트 프롬프트 부여
- Unbalanced Optimal Transport으로 로컬 Visual feature와 Global/Local 프롬프트를 정렬하여 데이터 이질성 극복



[2] Li, H., Diao, E., He, C., Li, Z., Zhao, L., Song, W., Zhu, J., Zhang, S., & Daumé III, H. (2024). Global and local prompts cooperation via optimal transport for federated learning. In CVPR.

FedOTP

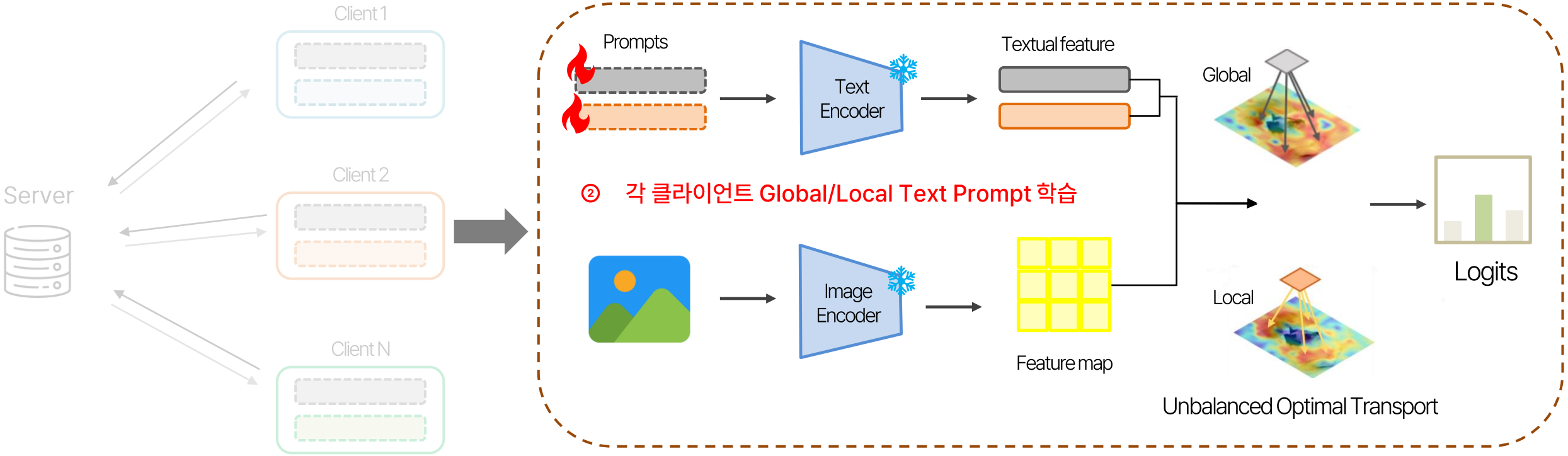
Vision Language Models-based Prompt Tuning for Federated Learning

: Global text prompt

: Local text prompt

❖ FedOTP: Global and local prompts cooperation via optimal transport for federated learning

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Global/Local 이중 텍스트 프롬프트 부여
- Unbalanced Optimal Transport으로 로컬 Visual feature와 Global/Local 프롬프트를 정렬하여 데이터 이질성 극복



[2] Li, H., Diao, E., He, C., Li, Z., Zhao, L., Song, W., Zhu, J., Zhang, S., & Daumé III, H. (2024). Global and local prompts cooperation via optimal transport for federated learning. In CVPR.

FedOTP

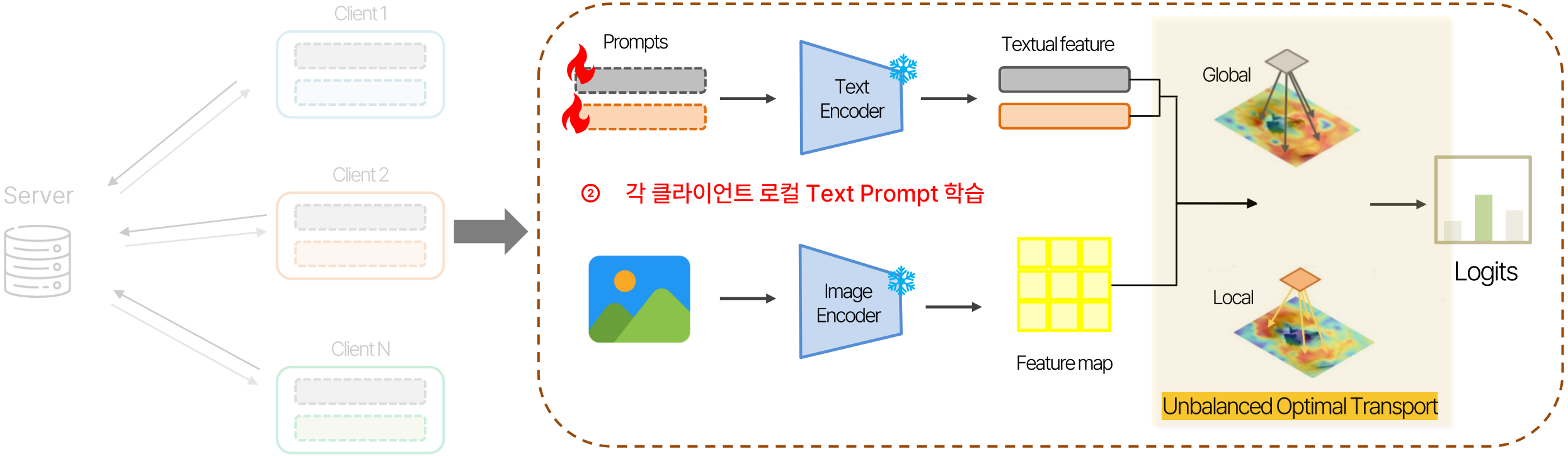
Vision Language Models-based Prompt Tuning for Federated Learning

: Global text prompt

: Local text prompt

❖ FedOTP: Global and local prompts cooperation via optimal transport for federated learning

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Global/Local 이중 텍스트 프롬프트 부여
- Unbalanced Optimal Transport으로 로컬 Visual feature와 Global/Local 프롬프트를 정렬하여 데이터 이질성 극복



[2] Li, H., Diao, E., He, C., Li, Z., Zhao, L., Song, W., Zhu, J., Zhang, S., & Daumé III, H. (2024). Global and local prompts cooperation via optimal transport for federated learning. In CVPR.

FedOTP

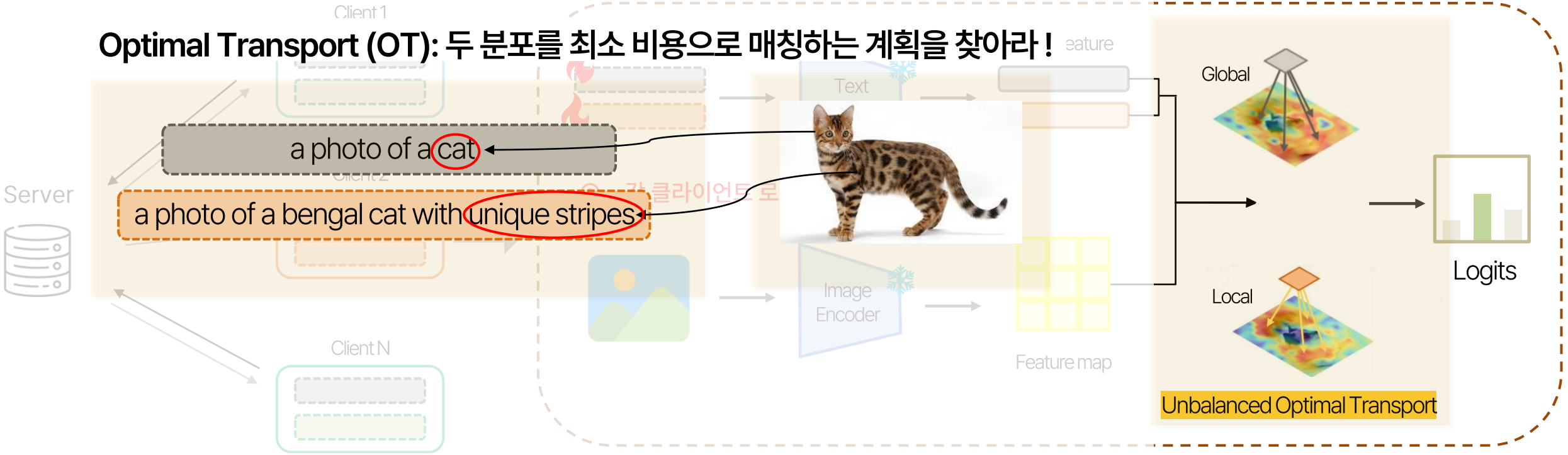
Vision Language Models-based Prompt Tuning for Federated Learning

: Global text prompt

: Local text prompt

❖ FedOTP: Global and local prompts cooperation via optimal transport for federated learning

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Global/Local 이중 텍스트 프롬프트 부여
- Unbalanced Optimal Transport으로 로컬 Visual feature와 Global/Local 프롬프트를 정렬하여 데이터 이질성 극복



[2] Li, H., Diao, E., He, C., Li, Z., Zhao, L., Song, W., Zhu, J., Zhang, S., & Daumé III, H. (2024). Global and local prompts cooperation via optimal transport for federated learning. In CVPR.

FedOTP

Vision Language Models-based Prompt Tuning for Federated Learning

: Global text prompt

: Local text prompt

❖ FedOTP: Global and local prompts cooperation via optimal transport for federated learning

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Global/Local 이중 텍스트 프롬프트 부여
- Unbalanced Optimal Transport으로 로컬 Visual feature와 Global/Local 프롬프트를 정렬하여 데이터 이질성 극복

Visual prompt → Text prompt → Transport plan

$$d_C(\alpha,\beta) = \min_{T\in U(\alpha,\beta)} \langle C, T\rangle,$$

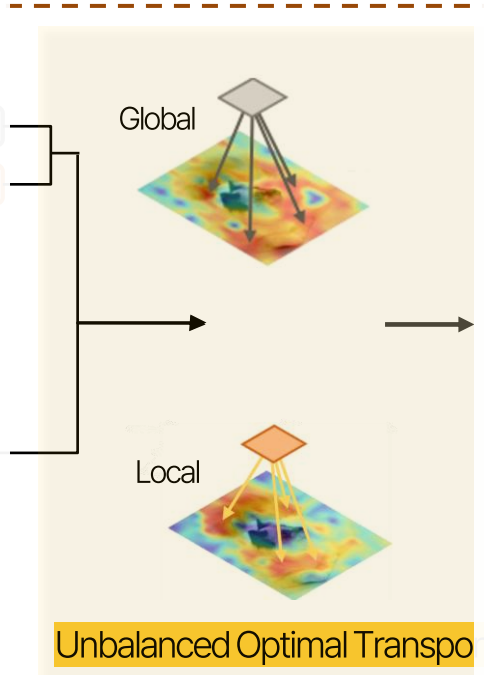
$$U(\alpha,\beta) = \left\{ T \in \mathbb{R}_+^{|\alpha|\times|\beta|} \mid T\mathbf{1}_{|\beta|} = \alpha,\ T^\top \mathbf{1}_{|\alpha|} = \beta \right\}$$

모든 visual patch가 반드시 Global/Local 텍스트 프롬프트에 모두 매칭되어야 함
→ 클래스와 무관한 patch도 강제 매칭

$$d_{C,k}(\alpha,\beta) = \min_{T\in U(\alpha,\beta)} \langle C, T\rangle,$$

$$U(\alpha,\beta) = \left\{ T \in \mathbb{R}_+^{\mathbb{V}\times 2} \mid T\mathbf{1}_{\mathbf{2}} \leq \alpha,\ T^\top \mathbf{1}_{\mathbf{V}} = \beta \right\}$$

모든 visual patch가 Global/Local 텍스트 프롬프트에 반드시 매칭될 필요 없음
→ 클래스와 관련 있는 patch만 선택적으로 매칭



[2] Li, H., Diao, E., He, C., Li, Z., Zhao, L., Song, W., Zhu, J., Zhang, S., & Daumé III, H. (2024). Global and local prompts cooperation via optimal transport for federated learning. In CVPR.

FedOTP

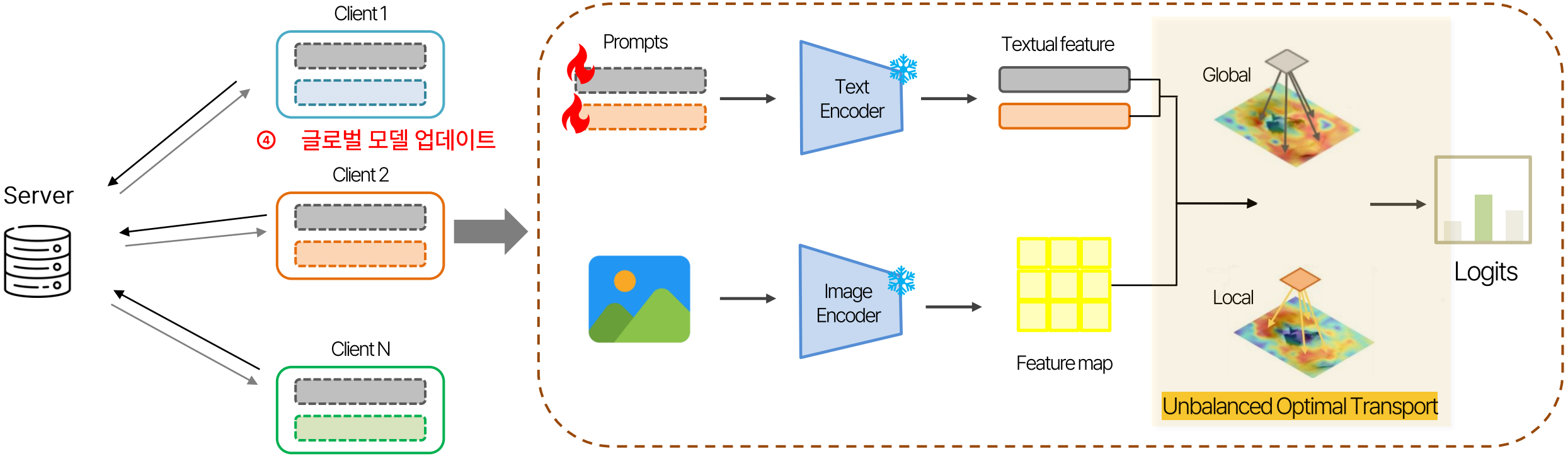
Vision Language Models-based Prompt Tuning for Federated Learning

: Global text prompt

: Local text prompt

❖ FedOTP: Global and local prompts cooperation via optimal transport for federated learning

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Global/Local 이중 텍스트 프롬프트 부여
- Unbalanced Optimal Transport으로 로컬 Visual feature와 Global/Local 프롬프트를 정렬하여 데이터 이질성 극복



[2] Li, H., Diao, E., He, C., Li, Z., Zhao, L., Song, W., Zhu, J., Zhang, S., & Daumé III, H. (2024). Global and local prompts cooperation via optimal transport for federated learning. In CVPR.

FedOTP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ Experiments

- 극심한 Label shift(클라이언트 간 클래스 중복x)에서 기존 방법들과 성능 비교
- FedOTP는 모든 데이터셋에서 SOTA 달성, PromptFL 대비 평균 +19%p 이상 성능 향상

Methods	Food101	DTD	Caltech101	Flowers102	OxfordPets
Local Training					
Zero-Shot CLIP [60]	75.27±0.05	40.21±0.12	85.14±0.24	62.17±0.12	84.47±0.10
CoOp [78]	82.54±2.42	82.69±0.63	90.41±0.44	88.23±0.76	94.52±1.30
Prompt-based Federated Learning					
PromptFL [27]	74.81±0.64	50.46±0.54	87.90±0.54	73.68±1.58	88.17±1.18
PromptFL+FT [24]	77.16±1.56	53.74±1.36	89.70±0.25	72.31±0.91	91.23±0.50
PromptFL+FedProx [42]	73.96±0.75	50.89±0.71	87.80±1.10	74.14±0.65	87.25±1.48
PromptFL+FedPer [1]	71.29±1.87	50.23±0.82	86.72±1.45	72.11±1.35	89.50±1.62
PromptFL+FedAMP [32]	74.48±1.71	47.16±0.92	87.31±1.60	69.10±0.13	80.21±0.44
pFedPrompt [26]	92.26±1.34	77.14±0.09	96.54±1.31	86.46±0.15	91.84±0.41
FedOTP (Ours)	**92.73±0.15**	**87.67±0.70**	**97.02±0.36**	**96.23±0.44**	**98.82±0.11**

[2] Li, H., Diao, E., He, C., Li, Z., Zhao, L., Song, W., Zhu, J., Zhang, S., & Daumé III, H. (2024). Global and local prompts cooperation via optimal transport for federated learning. In CVPR.

FedOTP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ Experiments

낮을 수록 이질성이 높음

- Feature(각 클라이언트에게 한 개의 도메인 부여) + Label shift(Dirichlet distribution, α = 0.1)에서 기존 방법들과 성능 비교
- FedOTP는 대부분의 도메인에서 높은 성능을 보이며, 평균 성능에서도 가장 우수한 결과를 보임

Datasets	DomainNet						
Domains	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Avg.
Local Training							
Zero-Shot CLIP [60]	8.72±1.73	12.48±3.78	8.53±4.32	9.31±0.69	9.13±2.55	11.96±2.80	10.02±2.65
CoOp [78]	44.40±14.89	45.68±16.53	**47.21±18.20**	**41.13±20.62**	48.02±24.49	39.47±5.68	44.32±16.74
Prompt-based Federated Learning							
PromptFL [27]	9.31±6.53	12.58±9.91	8.23± 8.47	14.79±12.07	9.37±10.82	7.48±11.32	10.29±10.35
PromptFL+FedProx [42]	9.84±6.60	11.16±11.17	10.64±6.79	13.40±16.09	9.39±7.69	6.78±11.76	10.20±10.99
FedOTP (Ours)	**46.14±6.53**	**60.14±18.23**	45.2±16.86	38.66±7.60	**49.30±17.80**	**49.02±24.22**	**48.08±15.21**

- Unbalanced OT 기반 FedOTP가 Similarity Averaging 및 Classical OT보다 전반적으로 가장 높은 성능을 보임

Datasets	Food101		DTD		Caltech101		Flowers102		OxfordPets	
Number of shots	2	8	2	8	2	8	2	8	2	8
FedOTP (Similarity Averaging)	83.38±0.54	87.59±1.05	81.01±0.23	88.17±0.73	92.68±0.44	96.73±0.29	91.73±0.68	97.09±0.18	96.23±0.25	98.34±0.15
FedOTP (Classical OT)	88.07±0.63	89.77±0.62	81.42±0.99	88.43±0.45	93.17±0.68	96.80±0.23	92.84±1.34	97.07±0.25	96.55±0.26	98.51±0.27
FedOTP (Unbalanced OT)	**89.12±0.28**	**92.94±0.18**	**85.50±0.35**	**90.25±0.74**	**95.05±0.49**	**97.34±0.18**	**93.96±0.48**	**98.23±0.32**	**97.73±0.57**	**99.02±0.38**

[2] Li, H., Diao, E., He, C., Li, Z., Zhao, L., Song, W., Zhu, J., Zhang, S., & Daumé III, H. (2024). Global and local prompts cooperation via optimal transport for federated learning. In CVPR.

FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

- 2025년 NeurIPS, 인용 수 2회
- 기존 FL 프롬프트 학습 방법들은 단일 프롬프트에 의존해 데이터 이질성 환경에서 불안정한 집계 발생
- Multi-group Text-Visual 이중 프롬프트 + Diversity Loss + Dynamic Aggregation으로 개인화와 일반화를 동시에 달성

FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

Weihao Bo[1], **Yanpeng Sun**[2]*, **Yu Wang**[3], **Xinyu Zhang**[4], **Zechao Li**[1]
[1]Nanjing University of Science and Technology
[2]National University of Singapore
[3]Baidu VIS
[4]University of Auckland

Abstract

In this paper, we introduce **FedMGP**, a new paradigm for personalized federated prompt learning in vision-language models (VLMs). Existing federated prompt learning (FPL) methods often rely on a single, text-only prompt representation, which leads to client-specific overfitting and unstable aggregation under heterogeneous data distributions. Toward this end, FedMGP equips each client with *multiple groups* of paired textual and visual prompts, enabling the model to capture diverse, fine-grained semantic and instance-level cues. A diversity loss is introduced to drive each prompt group to specialize in distinct and complementary se-

[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

- 2025년 NeurIPS, 인용 수 2회
- 기존 FL 프롬프트 학습 방법들은 단일 프롬프트에 의존해 데이터 이질성 환경에서 불안정한 집계 발생
- Multi-group Text-Visual 이중 프롬프트 + Diversity Loss + Dynamic Aggregation으로 개인화와 일반화를 동시에 달성

FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

Weihao Bo[1], **Yanpeng Sun**[2]*, **Yu Wang**[3], **Xinyu Zhang**[4], **Zechao Li**[1]
[1]Nanjing University of Science and Technology
[2]National University of Singapore
[3]Baidu VIS
[4]University of Auckland

Abstract

In this paper, we introduce **FedMGP**, a new paradigm for personalized federated prompt learning in vision-language models (VLMs). Existing federated prompt learning (FPL) methods often rely on a single, text-only prompt representation, which leads to client-specific overfitting and unstable aggregation under heterogeneous data distributions. Toward this end, FedMGP equips each client with *multiple groups* of paired textual and visual prompts, enabling the model to capture diverse, fine-grained semantic and instance-level cues. A diversity loss is introduced to drive each prompt group to specialize in distinct and complementary se-

[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

- 2025년 NeurIPS, 인용 수 2회
- 기존 FL 프롬프트 학습 방법들은 단일 프롬프트에 의존해 데이터 이질성 환경에서 불안정한 집계 발생
- Multi-group Text-Visual 이중 프롬프트 + Diversity Loss + Dynamic Aggregation으로 개인화와 일반화를 동시에 달성

FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

Weihao Bo[1], **Yanpeng Sun**[2]*, **Yu Wang**[3], **Xinyu Zhang**[4], **Zechao Li**[1]
[1]Nanjing University of Science and Technology
[2]National University of Singapore
[3]Baidu VIS
[4]University of Auckland

Abstract

In this paper, we introduce **FedMGP**, a new paradigm for personalized federated prompt learning in vision-language models (VLMs). Existing federated prompt learning (FPL) methods often rely on a single, text-only prompt representation, which leads to client-specific overfitting and unstable aggregation under heterogeneous data distributions. Toward this end, FedMGP equips each client with *multiple groups* of paired textual and visual prompts, enabling the model to capture diverse, fine-grained semantic and instance-level cues. A diversity loss is introduced to drive each prompt group to specialize in distinct and complementary se-

[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

FedMGP

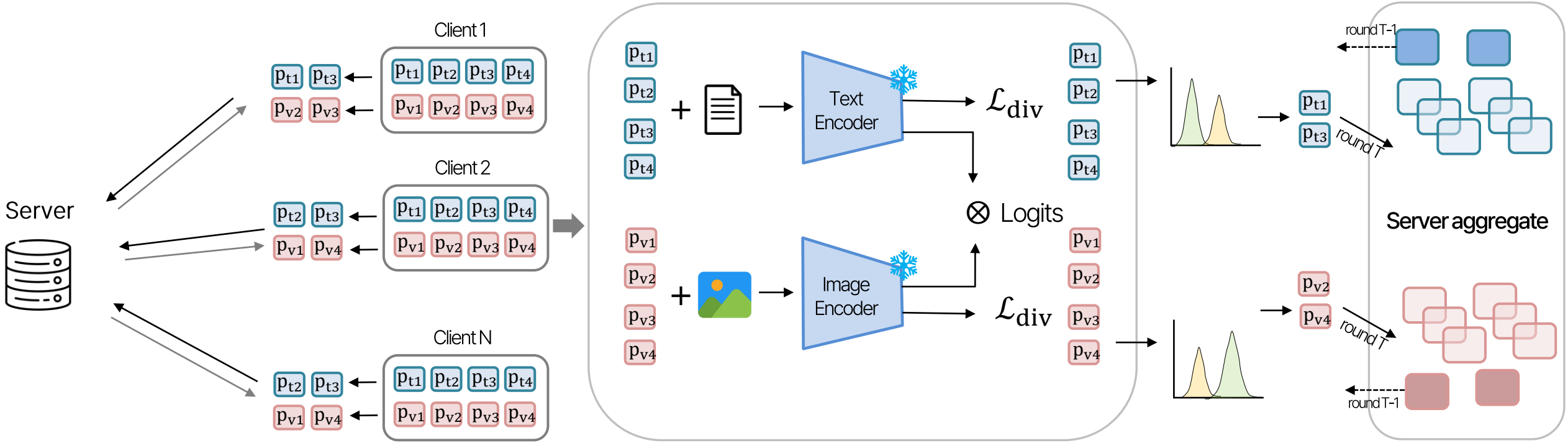
Vision Language Models-based Prompt Tuning for Federated Learning

: Text prompt

: Visual prompt

❖ FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Multi-group Text-Visual 이중 프롬프트 부여
- Diversity Loss으로 각 프롬프트 그룹이 서로 다른 의미적 특징에 집중 되도록 학습
- Dynamic Aggregation으로 유사한 클라이언트끼리 선택적 집계



[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

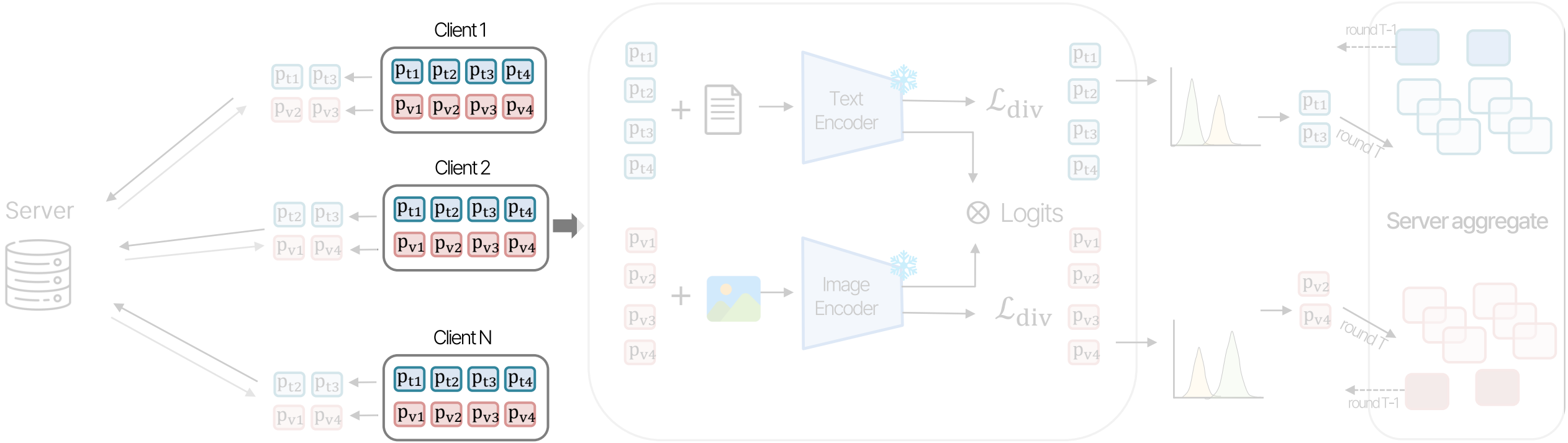
: Text prompt
: Visual prompt

FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Multi-group Text-Visual 이중 프롬프트 부여
- Diversity Loss으로 각 프롬프트 그룹이 서로 다른 의미적 특징에 집중 되도록 학습
- Dynamic Aggregation으로 유사한 클라이언트끼리 선택적 집계



[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

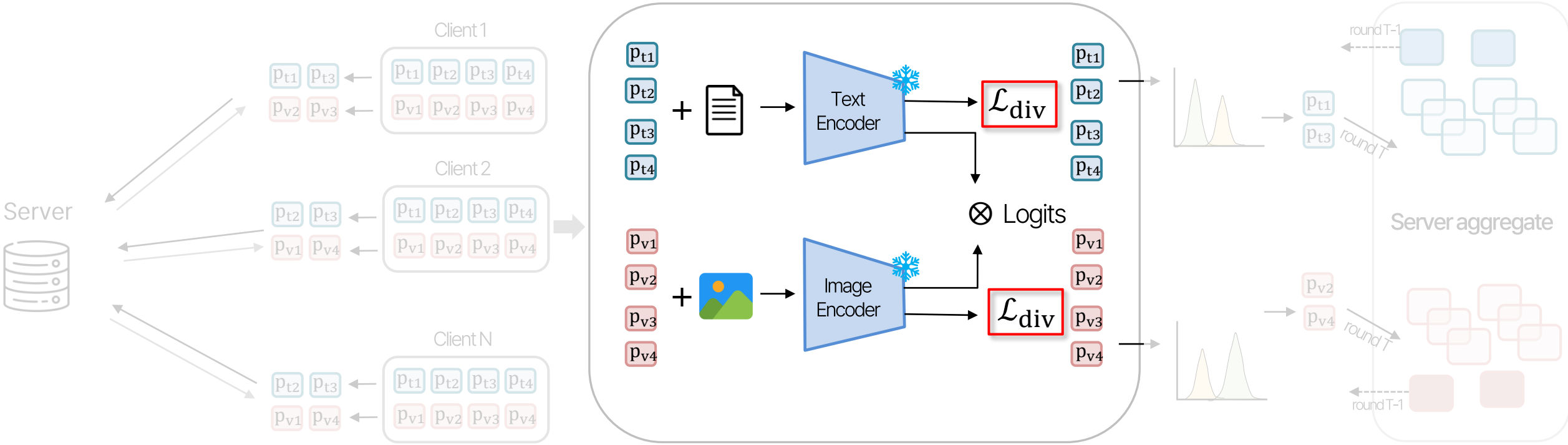
: Text prompt
: Visual prompt

FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Multi-group Text-Visual 이중 프롬프트 부여
- Diversity Loss으로 각 프롬프트 그룹이 서로 다른 의미적 특징에 집중 되도록 학습



[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

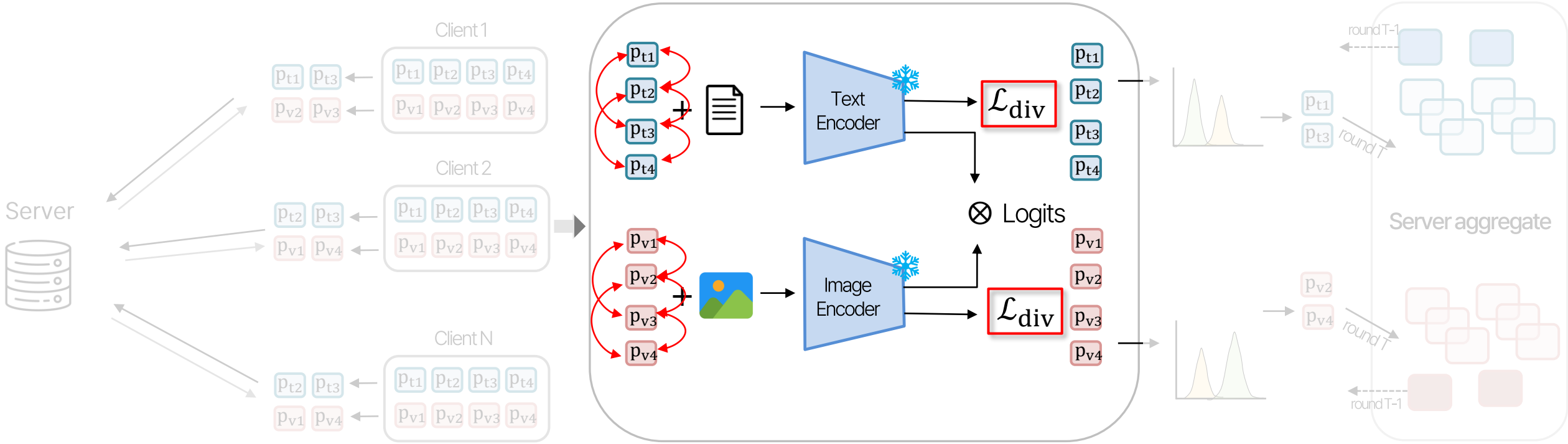
: Text prompt
: Visual prompt

FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Multi-group Text-Visual 이중 프롬프트 부여
- Diversity Loss으로 각 프롬프트 그룹이 서로 다른 의미적 특징에 집중 되도록 학습



[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Multi-group Text-Visual 이중 프롬프트 부여
- Diversity Loss으로 각 프롬프트 그룹이 서로 다른 의미적 특징에 집중 되도록 학습

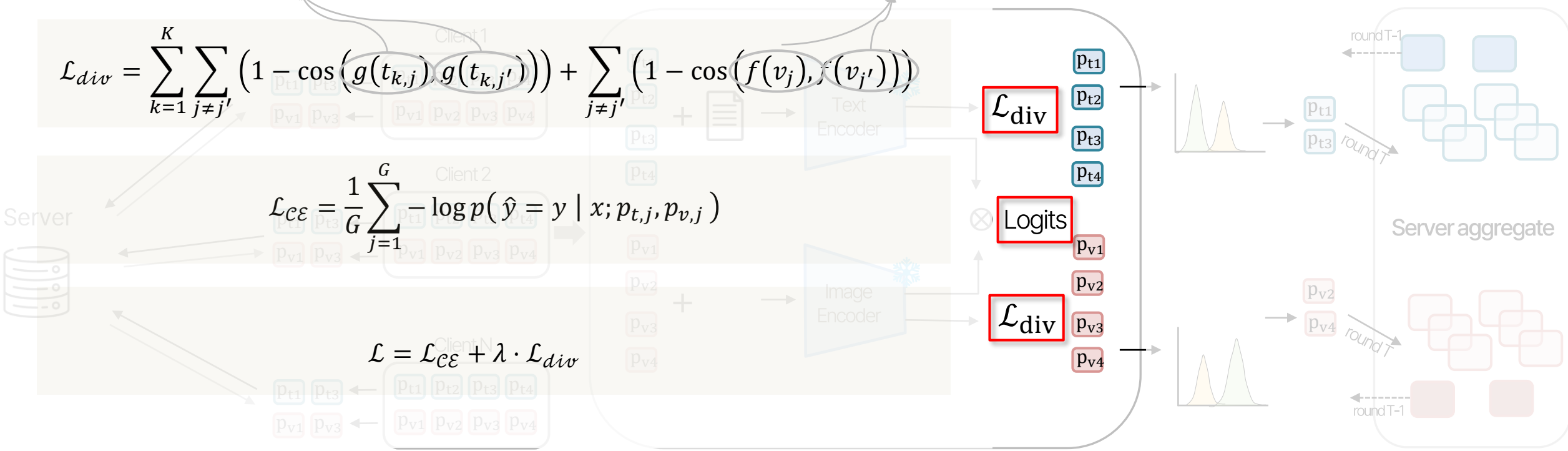
다른 텍스트 프롬프트 그룹끼리 서로 다른 부분에 집중 하도록 !

다른 비쥬얼 프롬프트 그룹끼리 서로 다른 부분에 집중 하도록 !

$$\mathcal{L}_{div} = \sum_{k=1}^{K} \sum_{j \neq j'} \left(1 - \cos\left(g(t_{k,j}), g(t_{k,j'})\right)\right) + \sum_{j \neq j'} \left(1 - \cos\left(f(v_j), f(v_{j'})\right)\right)$$

$$\mathcal{L}_{CE} = \frac{1}{G} \sum_{j=1}^{G} -\log p\left(\hat{y} = y \mid x; p_{t,j}, p_{v,j}\right)$$

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{div}$$



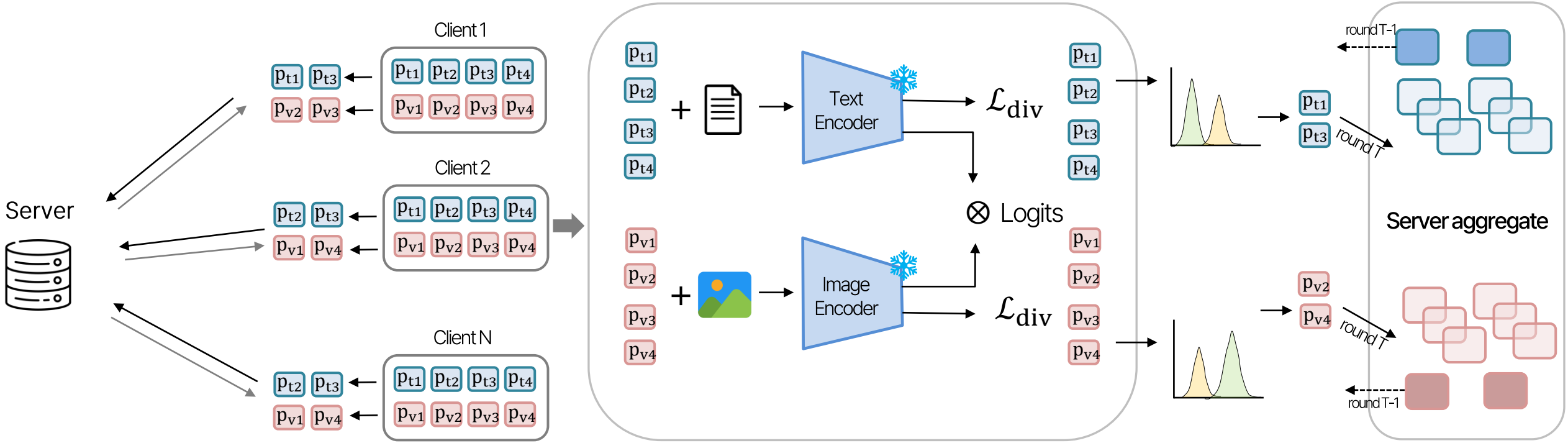
[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Multi-group Text-Visual 이중 프롬프트 부여
- Diversity Loss으로 각 프롬프트 그룹이 서로 다른 의미적 특징에 집중 되도록 학습
- Dynamic Aggregation으로 유사한 클라이언트끼리 선택적 집계



[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

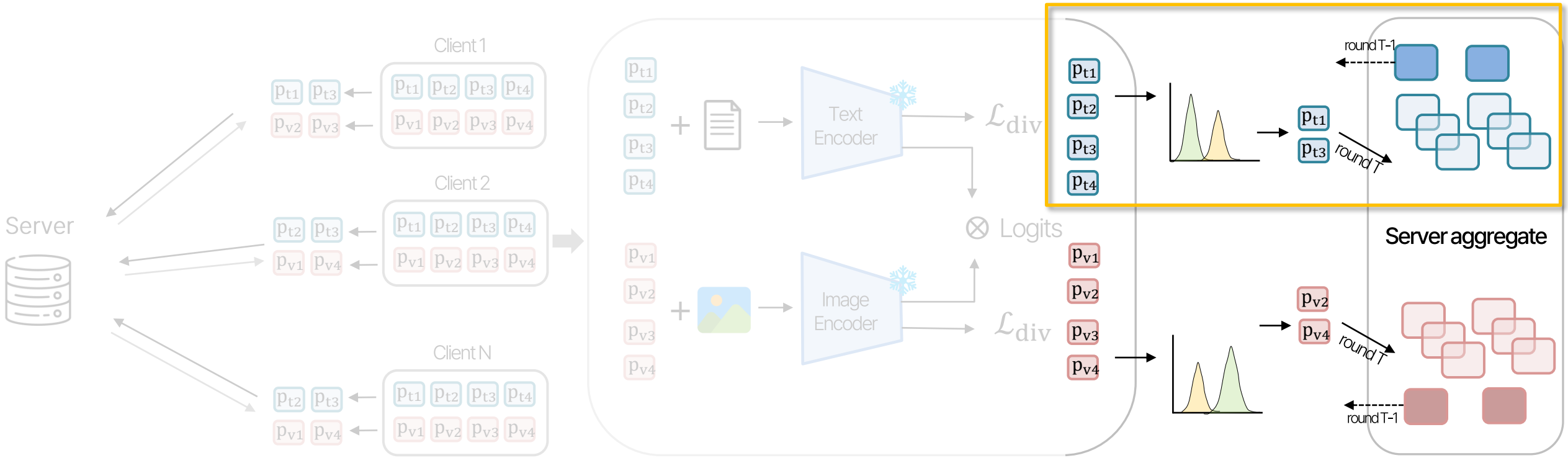
FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Multi-group Text-Visual 이중 프롬프트 부여
- Diversity Loss으로 각 프롬프트 그룹이 서로 다른 의미적 특징에 집중 되도록 학습
- Dynamic Aggregation으로 유사한 클라이언트끼리 선택적 집계

텍스트 / 비주얼 프롬프트 각각
이전 라운드 글로벌 프롬프트와의 유사도를 계산



[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

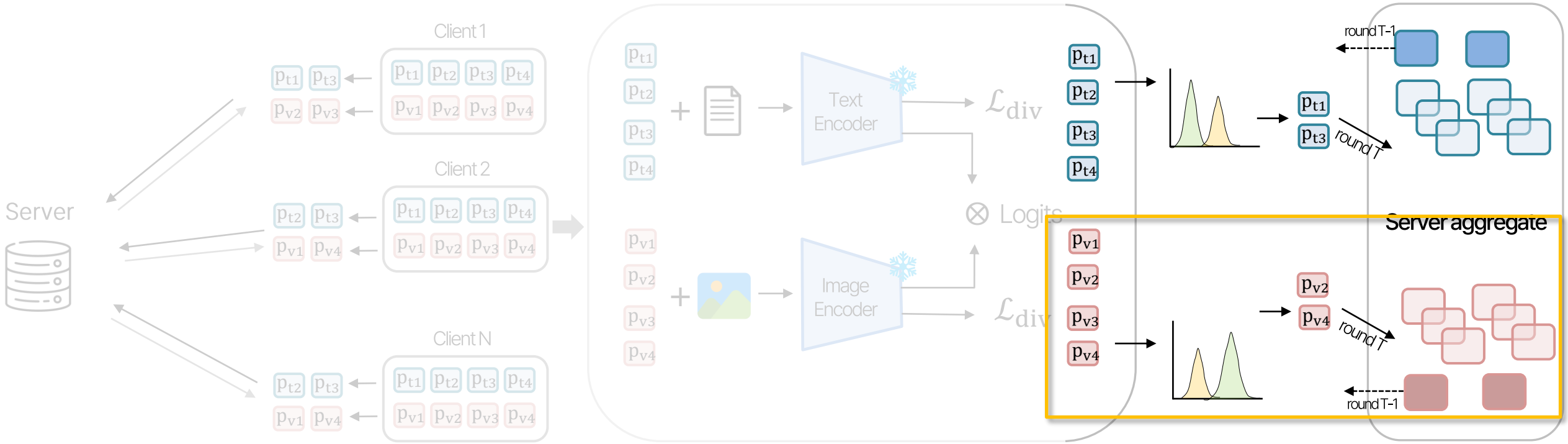
FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Multi-group Text-Visual 이중 프롬프트 부여
- Diversity Loss으로 각 프롬프트 그룹이 서로 다른 의미적 특징에 집중 되도록 학습
- Dynamic Aggregation으로 유사한 클라이언트끼리 선택적 집계

텍스트 / 비주얼 프롬프트 각각
이전 라운드 글로벌 프롬프트와의 유사도를 계산



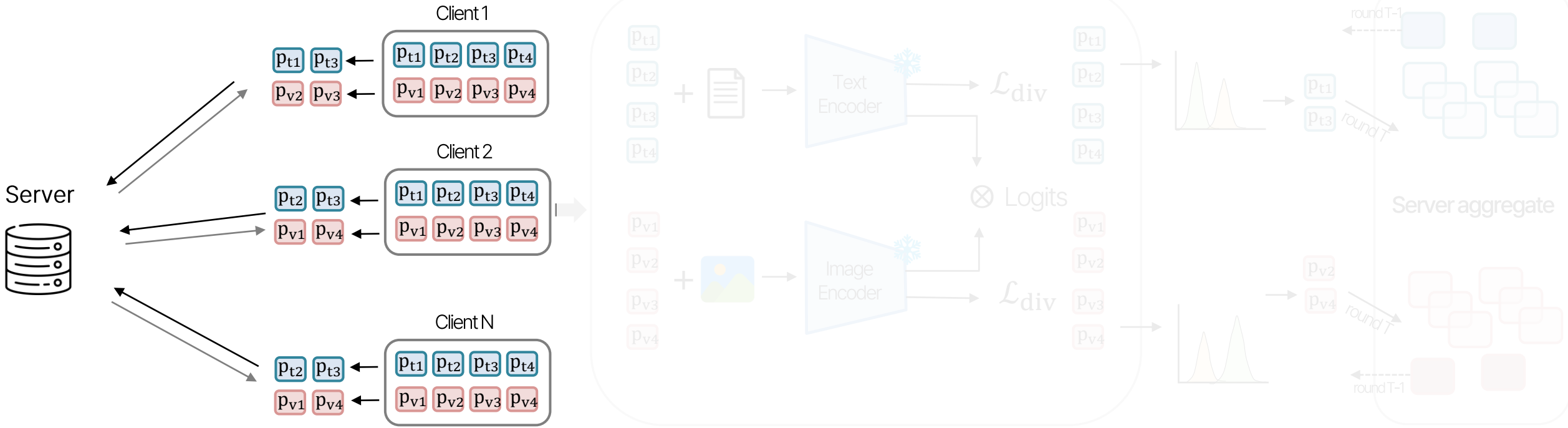
[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Multi-group Text-Visual 이중 프롬프트 부여
- Diversity Loss으로 각 프롬프트 그룹이 서로 다른 의미적 특징에 집중 되도록 학습
- Dynamic Aggregation으로 유사한 클라이언트끼리 선택적 집계



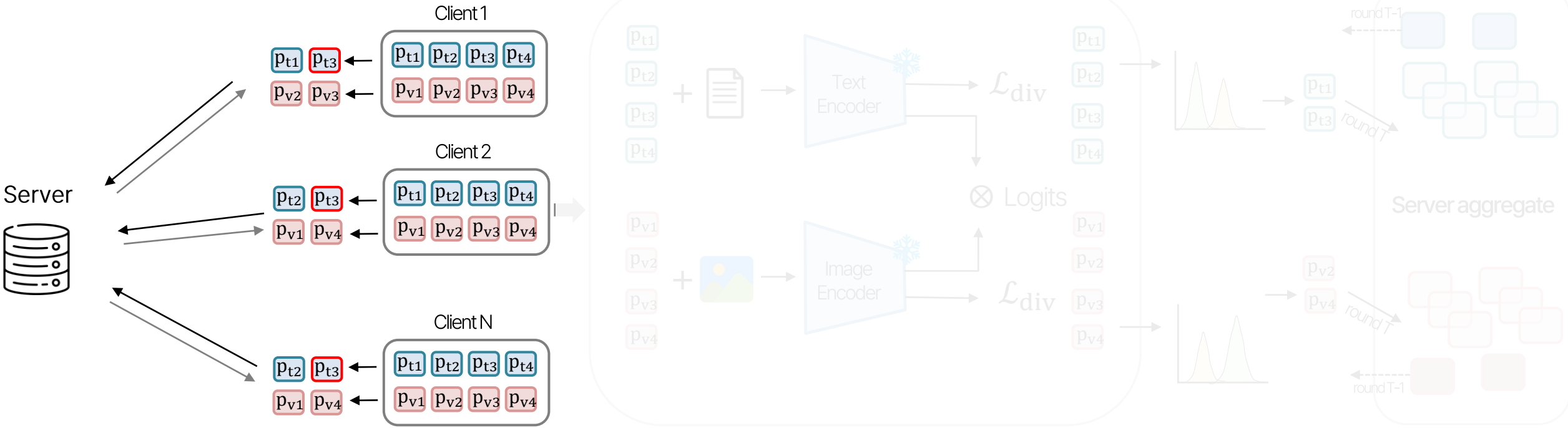
[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Multi-group Text-Visual 이중 프롬프트 부여
- Diversity Loss으로 각 프롬프트 그룹이 서로 다른 의미적 특징에 집중 되도록 학습
- Dynamic Aggregation으로 유사한 클라이언트끼리 선택적 집계



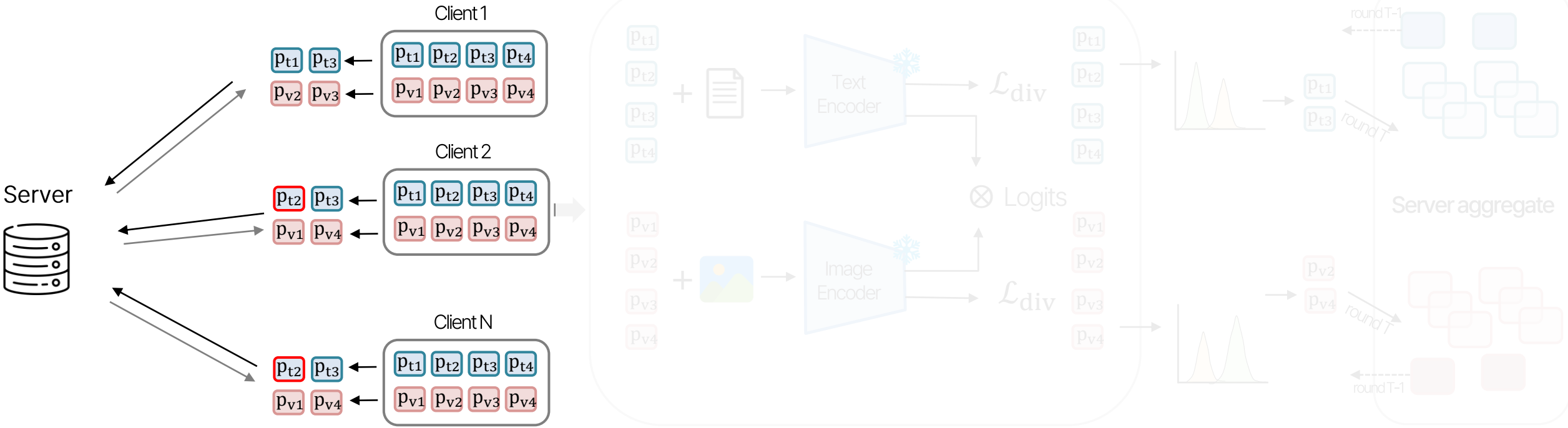
[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Multi-group Text-Visual 이중 프롬프트 부여
- Diversity Loss으로 각 프롬프트 그룹이 서로 다른 의미적 특징에 집중 되도록 학습
- Dynamic Aggregation으로 유사한 클라이언트끼리 선택적 집계



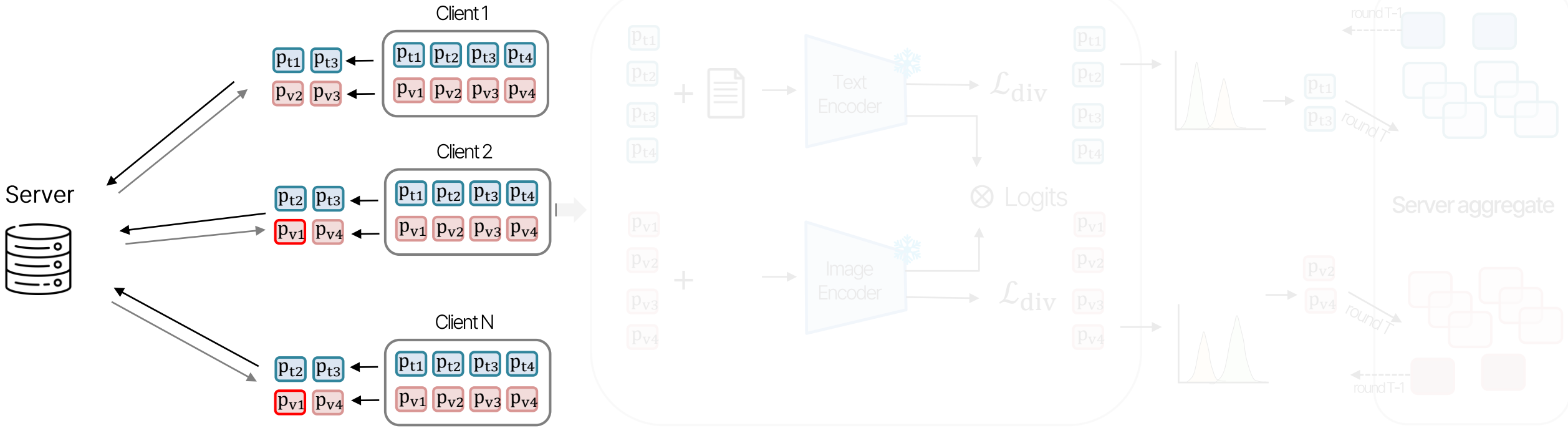
[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Multi-group Text-Visual 이중 프롬프트 부여
- Diversity Loss으로 각 프롬프트 그룹이 서로 다른 의미적 특징에 집중 되도록 학습
- Dynamic Aggregation으로 유사한 클라이언트끼리 선택적 집계



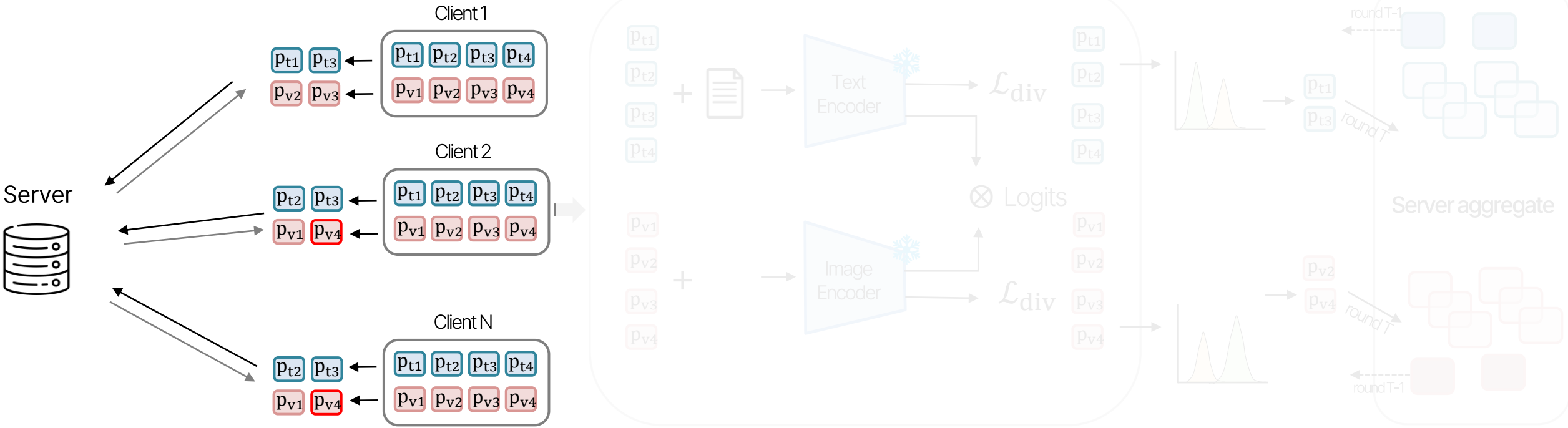
[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Multi-group Text-Visual 이중 프롬프트 부여
- Diversity Loss으로 각 프롬프트 그룹이 서로 다른 의미적 특징에 집중 되도록 학습
- Dynamic Aggregation으로 유사한 클라이언트끼리 선택적 집계



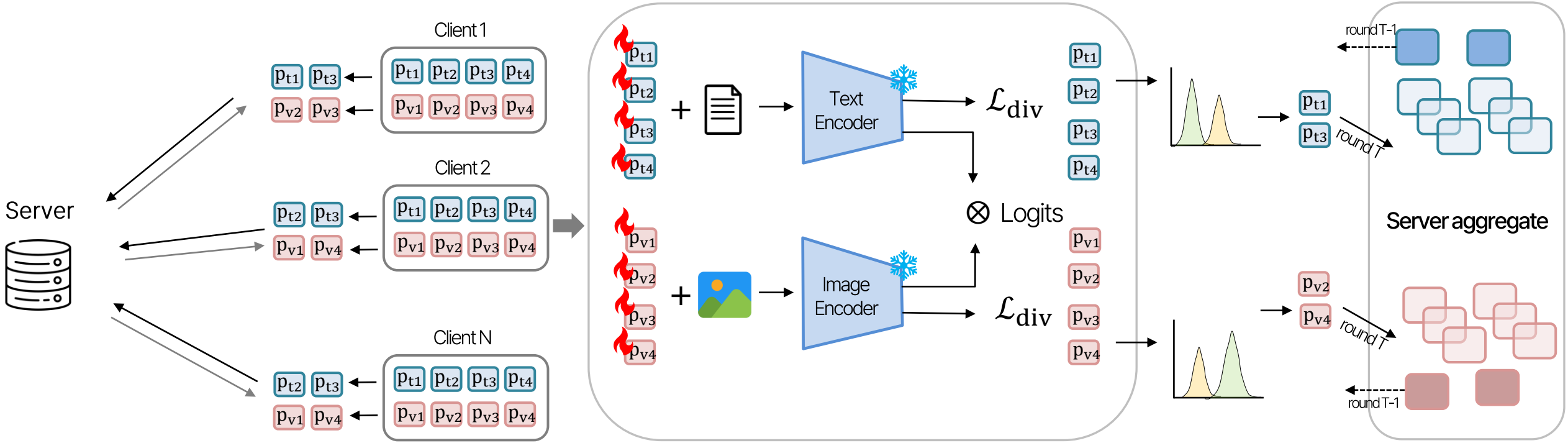
[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ FedMGP: Personalized Federated Learning with Multi-Group Text-Visual Prompts

- VLM(CLIP)의 Text/Image Encoder를 동결하고, 각 클라이언트에 Multi-group Text-Visual 이중 프롬프트 부여
- Diversity Loss으로 각 프롬프트 그룹이 서로 다른 의미적 특징에 집중 되도록 학습
- Dynamic Aggregation으로 유사한 클라이언트끼리 선택적 집계



[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

❖ Experiments

- 9개 데이터셋 기준 Local / Base / Novel / CM 지표로 기존 방법들과 성능 비교
- FedMGP는 평균 및 대부분의 데이터셋에서 SOTA 달성, 특히 CM 지표 81.85로 최고 성능

(a) Average over 9 datasets.

Methods	Local	Base	Novel	CM
PromptFL [17]	71.19	71.70	71.46	71.31
FedOTP [27]	92.53	16.84	31.66	57.10
FedTPG [41]	71.62	71.91	68.32	70.66
FedPGP [9]	84.32	72.45	68.97	77.42
PromptFolio [37]	96.02	39.75	51.02	70.29
FedMGP	93.17	68.49	72.99	**81.85**

(b) OxfordPets.

Methods	Local	Base	Novel	CM
PromptFL [17]	89.77	90.01	97.20	91.62
FedOTP [27]	100.00	26.68	57.16	68.19
FedTPG [41]	94.24	94.31	96.64	94.85
FedPGP [9]	96.20	95.01	96.89	96.07
PromptFolio [37]	99.90	66.23	83.38	86.86
FedMGP	97.15	93.83	97.04	**96.28**

(g) UCF101.

Methods	Local	Base	Novel	CM
PromptFL [17]	77.08	76.94	70.36	75.29
FedOTP [27]	92.39	16.33	19.07	54.99
FedTPG [41]	76.22	75.96	72.09	75.10
FedPGP [9]	82.61	71.78	68.45	76.34
PromptFolio [37]	96.15	31.94	42.00	66.22
FedMGP	92.69	68.38	72.86	**81.62**

(h) SUN397.

Methods	Local	Base	Novel	CM
PromptFL [17]	76.25	76.20	75.68	76.09
FedOTP [27]	93.40	11.38	19.11	53.83
FedTPG [41]	73.72	73.71	75.17	74.08
FedPGP [9]	89.43	66.51	67.43	78.20
PromptFolio [37]	95.18	32.89	44.47	66.50
FedMGP	91.83	68.51	72.20	**81.07**

(c) Flowers102.

Methods	Local	Base	Novel	CM
PromptFL [17]	70.33	71.79	75.39	71.94
FedOTP [27]	99.73	13.06	21.51	57.99
FedTPG [41]	79.43	78.92	73.26	77.71
FedPGP [9]	91.83	80.22	68.46	82.85
PromptFolio [37]	99.82	27.36	39.34	66.05
FedMGP	98.41	70.06	74.71	**85.36**

(d) DTD.

Methods	Local	Base	Novel	CM
PromptFL [17]	55.32	57.06	44.32	52.60
FedOTP [27]	96.44	20.06	41.23	61.71
FedTPG [41]	56.90	59.26	40.46	52.49
FedPGP [9]	78.47	67.22	50.93	68.21
PromptFolio [37]	97.18	26.53	37.39	64.11
FedMGP	92.87	53.60	55.62	**73.73**

(i) Stanford Cars.

Methods	Local	Base	Novel	CM
PromptFL [17]	62.98	63.14	69.87	64.66
FedOTP [27]	91.06	9.32	10.62	50.49
FedTPG [41]	65.50	65.47	69.10	66.37
FedPGP [9]	85.37	57.63	60.19	72.13
PromptFolio [37]	96.44	29.43	46.77	66.28
FedMGP	92.61	56.48	71.19	**77.80**

(j) FGVC Aircraft.

Methods	Local	Base	Novel	CM
PromptFL [17]	25.03	25.03	24.48	24.89
FedOTP [27]	64.34	7.27	8.12	36.01
FedTPG [41]	12.00	12.00	4.50	9.27
FedPGP [9]	47.59	25.89	22.89	35.94
PromptFolio [37]	82.50	12.29	17.09	48.40
FedMGP	78.46	21.03	30.15	**51.62**

(e) Caltech101.

Methods	Local	Base	Novel	CM
PromptFL [17]	94.16	95.35	94.98	94.66
FedOTP [27]	99.96	28.28	62.26	69.43
FedTPG [41]	96.17	97.16	91.92	95.32
FedPGP [9]	96.91	97.35	94.37	96.37
PromptFolio [37]	99.79	73.69	81.10	88.50
FedMGP	99.47	96.02	93.61	**97.13**

(f) Food101.

Methods	Local	Base	Novel	CM
PromptFL [17]	89.75	89.79	90.86	90.04
FedOTP [27]	95.44	19.16	45.89	61.24
FedTPG [41]	90.36	90.42	91.78	90.73
FedPGP [9]	90.51	90.48	91.12	90.65
PromptFolio [37]	97.24	57.40	67.64	79.67
FedMGP	95.08	88.47	89.53	**92.04**

[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

Class 종류: 개, 고양이, 사자, 돼지

[Train] Client1: 개
Client2: 고양이

[Test] Client1: 개 → Local
Client1: 고양이 → Base
Client1: 사자 → Novel

$$HM = \frac{2 \times Base \times Novel}{Base + Novel}$$

$$CM = \frac{Local + HM}{2}$$

❖ Experiments

- 9개 데이터셋 기준 Local / Base / Novel / CM 지표로 기존 방법들과 성능 비교
- FedMGP는 평균 및 대부분의 데이터셋에서 SOTA 달성, 특히 CM 지표 81.85로 최고 성능

(a) Average over 9 datasets.

Methods	Local	Base	Novel	CM
PromptFL [17]	71.19	71.70	71.46	71.31
FedOTP [27]	92.53	16.84	31.66	57.10
FedTPG [41]	71.62	71.91	68.32	70.66
FedPGP [9]	84.32	72.45	68.97	77.42
PromptFolio [37]	96.02	39.75	51.02	70.29
FedMGP	93.17	68.49	72.99	**81.85**

(b) OxfordPets.

Methods	Local	Base	Novel	CM
PromptFL [17]	89.77	90.01	97.20	91.62
FedOTP [27]	100.00	26.68	57.16	68.19
FedTPG [41]	94.24	94.31	96.64	94.85
FedPGP [9]	96.20	95.01	96.89	96.07
PromptFolio [37]	99.90	66.23	83.38	86.86
FedMGP	97.15	93.83	97.04	**96.28**

(g) UCF101.

Methods	Local	Base	Novel	CM
PromptFL [17]	77.08	76.94	70.36	75.29
FedOTP [27]	92.39	16.33	19.07	54.99
FedTPG [41]	76.22	75.96	72.09	75.10
FedPGP [9]	82.61	71.78	68.45	76.34
PromptFolio [37]	96.15	31.94	42.00	66.22
FedMGP	92.69	68.38	72.86	**81.62**

(h) SUN397.

Methods	Local	Base	Novel	CM
PromptFL [17]	76.25	76.20	75.68	76.09
FedOTP [27]	93.40	11.38	19.11	53.83
FedTPG [41]	73.72	73.71	75.17	74.08
FedPGP [9]	89.43	66.51	67.43	78.20
PromptFolio [37]	95.18	32.89	44.47	66.50
FedMGP	91.83	68.51	72.20	**81.07**

(c) Flowers102.

Methods	Local	Base	Novel	CM
PromptFL [17]	70.33	71.79	75.39	71.94
FedOTP [27]	99.73	13.06	21.51	57.99
FedTPG [41]	79.43	78.92	73.26	77.71
FedPGP [9]	91.83	80.22	68.46	82.85
PromptFolio [37]	99.82	27.36	39.34	66.05
FedMGP	98.41	70.06	74.71	**85.36**

(d) DTD.

Methods	Local	Base	Novel	CM
PromptFL [17]	55.32	57.06	44.32	52.60
FedOTP [27]	96.44	20.06	41.23	61.71
FedTPG [41]	56.90	59.26	40.46	52.49
FedPGP [9]	78.47	67.22	50.93	68.21
PromptFolio [37]	97.18	26.53	37.39	64.11
FedMGP	92.87	53.60	55.62	**73.73**

(i) Stanford Cars.

Methods	Local	Base	Novel	CM
PromptFL [17]	62.98	63.14	69.87	64.66
FedOTP [27]	91.06	9.32	10.62	50.49
FedTPG [41]	65.50	65.47	69.10	66.37
FedPGP [9]	85.37	57.63	60.19	72.13
PromptFolio [37]	96.44	29.43	46.77	66.28
FedMGP	92.61	56.48	71.19	**77.80**

(j) FGVC Aircraft.

Methods	Local	Base	Novel	CM
PromptFL [17]	25.03	25.03	24.48	24.89
FedOTP [27]	64.34	7.27	8.12	36.01
FedTPG [41]	12.00	12.00	4.50	9.27
FedPGP [9]	47.59	25.89	22.89	35.94
PromptFolio [37]	82.50	12.29	17.09	48.40
FedMGP	78.46	21.03	30.15	**51.62**

(e) Caltech101.

Methods	Local	Base	Novel	CM
PromptFL [17]	94.16	95.35	94.98	94.66
FedOTP [27]	99.96	28.28	62.26	69.43
FedTPG [41]	96.17	97.16	91.92	95.32
FedPGP [9]	96.91	97.35	94.37	96.37
PromptFolio [37]	99.79	73.69	81.10	88.50
FedMGP	99.47	96.02	93.61	**97.13**

(f) Food101.

Methods	Local	Base	Novel	CM
PromptFL [17]	89.75	89.79	90.86	90.04
FedOTP [27]	95.44	19.16	45.89	61.24
FedTPG [41]	90.36	90.42	91.78	90.73
FedPGP [9]	90.51	90.48	91.12	90.65
PromptFolio [37]	97.24	57.40	67.64	79.67
FedMGP	95.08	88.47	89.53	**92.04**

[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

FedMGP

Vision Language Models-based Prompt Tuning for Federated Learning

Class 종류: 개, 고양이, 사자, 돼지

[Train] Client1: 개
Client2: 고양이

[Test] Client1: 개 → Local
Client1: 고양이 → Base
Client1: 사자 → Novel

$$HM = \frac{2 \times Base \times Novel}{Base + Novel}$$

$$CM = \frac{Local + HM}{2}$$

❖ Experiments

<Ablation study on prompt length>

Setting	Local	Base	Novel	CM
FedMGP (l=4)	97.18	72.49	72.17	84.75
FedMGP (l=8)	98.05	64.00	64.91	81.25
FedMGP (l=16)	97.62	57.47	61.56	78.53
FedMGP (l=2)	96.92	73.23	74.65	**85.43**

프롬프트가 길면 오히려 성능 하락

<Ablation study on Prompt Groups>

Setting	Local	Base	Novel	CM
FedMGP (m=4)	96.60	73.28	73.99	85.12
FedMGP (m=3)	89.95	77.68	74.48	83.00
FedMGP (m=2)	82.88	82.15	74.05	80.38
FedMGP (m=1)	78.85	79.68	70.67	76.88
FedMGP (m=5)	96.92	73.23	74.65	**85.43**

그룹 수가 적으면 다양성 부족

<Ablation study on Top-s>

Setting	Local	Base	Novel	CM
FedMGP (Top-s=1)	97.88	69.17	74.10	84.72
FedMGP (Topk-s=3)	92.93	76.88	74.85	84.39
FedMGP (Topk-s=4)	86.77	79.44	74.89	81.93
FedMGP (Topk-s=2)	96.92	73.23	74.65	**85.43**

너무 많은 그룹 선택 시 불필요한 정보 포함

<Ablation study on the impact of vision and text prompt>

Setting	Local	Base	Novel	CM
FedMGP (Vision Only)	75.94	76.48	72.92	75.30
FedMGP (Text Only)	95.23	73.60	73.80	84.46
FedMGP (Vision + Text)	96.92	73.23	74.65	**85.43**

Text Only 대비 Vision+Text 이중 프롬프트가 더 높음

< Ablation study on $\mathcal{L}_{div}$ >

Setting	Local	Base	Novel	CM
FedMGP (w/o $\mathcal{L}_{\text{div}}$)	94.53	72.97	72.48	83.63
FedMGP ($\mathcal{L}_{\text{div}}$=2)	95.78	74.88	74.98	85.35
FedMGP ($\mathcal{L}_{\text{div}}$=5)	96.35	73.50	74.31	85.13
FedMGP ($\mathcal{L}_{\text{div}}$=10)	95.78	73.09	74.35	84.75
FedMGP ($\mathcal{L}_{\text{div}}$=1)	96.92	73.23	74.65	**85.43**

Diversity Loss의 필요성 입증

[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

Conclusion

Vision Language Models-based Prompt Tuning for Federated Learning

- **How to apply VLM prompt tuning in Federated Learning?**

1. PromptFL(2024, IEEE Transactions on Mobile Computing) : FL 환경에서 텍스트 소수의 프롬프트만을 학습·집계하는 방법 최초 제안
   - VLM 전체 파라미터 대신 텍스트 프롬프트 벡터만 공유 → 통신 비용 0.01%로 절감
   - 단순 FedAvg 집계로 데이터 이질성 환경에서 개인화 전략 부재

2. FedOTP (2024, CVPR) : Global/Local 이중 텍스트 프롬프트 + Unbalanced OT로 데이터 이질성 극복
   - 클라이언트별 공통 지식과 개인화 정보를 분리하여 동시에 학습
   - 텍스트 프롬프트에만 국한, 비주얼 modality 활용 X

3. FedMGP (2025, NeurIPS) : Multi-group Text-Visual 이중 프롬프트로 텍스트를 넘어 비주얼까지 확장
   - Diversity Loss로 각 그룹이 서로 다른 의미적 특징에 집중
   - Dynamic Aggregation으로 유사한 클라이언트끼리 선택적 집계

[1] Guo, T., Guo, S., Wang, J., Tang, X., & Xu, W. (2023). PromptFL: Let federated participants cooperatively learn prompts instead of models – federated learning in age of foundation model. IEEE Transactions on Mobile Computing.
[2] Li, H., Diao, E., He, C., Li, Z., Zhao, L., Song, W., Zhu, J., Zhang, S., & Daumé III, H. (2024). Global and local prompts cooperation via optimal transport for federated learning. In CVPR.
[3] Bo, W., Sun, Y., Wang, Y., Zhang, X., & Li, Z. (2025). FedMGP: Personalized federated learning with multi-group text-visual prompts. In NeurIPS.

Thank you