Complexity-Aware LoRA Aggregation for Modality-Heterogeneous Federated Person Re-identification

Kunyang Lv[1], **Wenke Huang**[1], **Wenwen He**[1],
Bin Yang[1]*, **Bo Du**[1], **Mang Ye**[1]
[1]School of Computer Science, Wuhan University
{lvkunyang, yangbin_cv}@whu.edu.cn

Abstract

A unified cross-modal person re-identification targets robust cross-modal retrieval with queries from diverse modalities like visible, infrared, and text. Such framework requires multi-modal data and centralized training, raising data costs and privacy risks. Hence, we propose a modality-heterogeneous federated ReID task to learn a unified cross-modal framework from clients with different cross-modal data. However, it faces two challenges. First, varying inter-modality complexity across tasks leads to different retrieval patterns and parameter space. Second, different intra-modality complexity results in imbalanced contributions when merging LoRA modules. Existing federated learning methods use uniform merging and transmit full models, which hinders cross-modal knowledge integration and increases communication overhead. To address these issues, we adopt Low-Rank Adaptation (LoRA), a lightweight module for fine-tuning large models through low-rank updates, enabling clients to send compact parameters. On top of LoRA, we develop MIRAGE (Modality-heterogeneous Intelligent federated ReID Aggregation with Global Efficiency), adapting aggregation to inter- and intra-modality complexity. When merging visible LoRA, MIRAGE selects key ranks to protect retrieval patterns while retaining shared ranks for common ability. As for infrared and text LoRA, it reduces conflicts with an adaptive pruning method. In both strategies, task complexity is incorporated into aggregation to balance contributions. Experiments on cross-modal retrieval tasks demonstrate the superiority of MIRAGE. The code and dataset are available at project.

1 Introduction

Given a query image or detailed textual description, Person Re-identification (ReID) aims to identify the accurate pedestrian in a large collection of images or videos [65, 77, 46]. It is widely applied in security domains like intelligent surveillance and smart cities. Recently, ReID has developed rapidly, achieving remarkable performance in single [1, 76, 23] and cross-modality tasks [69, 21, 38, 61, 63].

Real-world sensing environments are diverse. Different modalities offer unique advantages under specific conditions, driving the development of cross-modal ReID. Existing cross-modal ReID methods match pedestrians across modalities to ensure performance. Under low-light circumstance, visible cameras degrade while infrared sensors remain reliable, making visible–infrared ReID learn shared spectral features [60, 47, 61, 59]. Text–image ReID is particularly useful when surveillance data is missing and clues rely on witness descriptions. It aligns semantic cues in text with visual

*Corresponding Author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

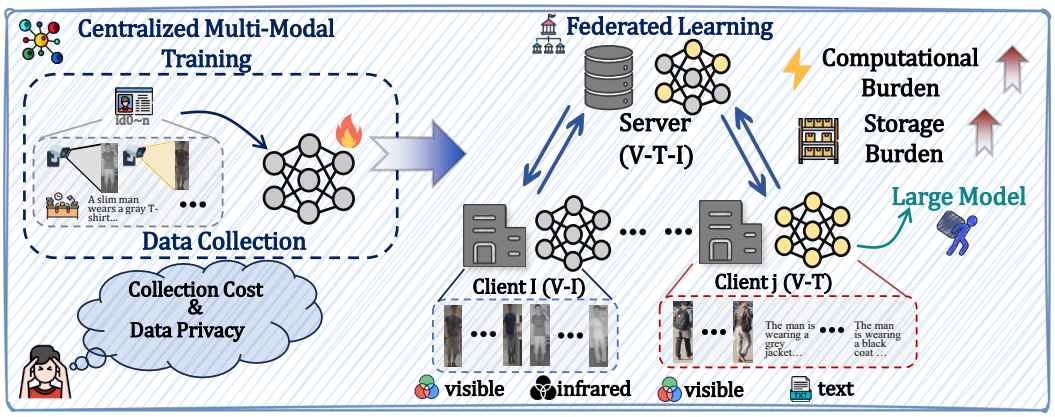


Figure 1: **Motivation.** The centralized multi-modal training results in high data cost and privacy leakage. Modality-heterogeneous federated learning supporting visible (V), infrared (I), and text (T) inputs transmits entire modal, which brings high communication overhead.

attributes [38, 20]. These approaches all target specific modality pairs. In reality, queries may come from diverse sources, *e.g.*, visible (V), infrared (I), and text (T), limiting the pairwise methods.

Handling diverse modality pairs in cross-modal ReID often requires separate models, causing significant computational and deployment overhead. Existing multi-modal ReID approaches [29, 71, 4] rely on multi-modal data and centralized training. However, constructing such datasets is costly and the centralized training raises critical privacy concerns. Federated learning (FL) [37, 41, 5, 73, 30, 54] provides an alternative. It enables multiple clients to jointly train a model without sharing raw data. To address these issues, we propose a modality-heterogeneous federated ReID framework where each client handles a unique cross-modal task (*i.e.* V–I, V–T, and I–T). These clients often employ strong encoders such as ResNet-50 [14, 65] or ViT [15, 7], leading to high communication overhead in conventional FL. Though this overhead limits scalability, LoRA [17], a parameter-efficient fine-tuning method, can greatly reduce communication costs while maintaining comparable performance. Motivated by this, we incorporate LoRA by injecting, for each client, two LoRA modules for its modalities (*e.g.* visual LoRA and textual LoRA in the V–T client). The server then merges LoRA parameters of the same modality from different clients.

Despite the advantage in preserving modality-specific knowledge, our scheme faces critical challenges. One obstacle arises from the variation in inter-modality complexity across different cross-modal ReID tasks. Visible images contain highly fine-grained details, making accurate recognition inherently difficult. In V–T and V–I tasks, visible inputs tend to produce distinct retrieval patterns, leading to minimal overlap in their visible LoRA parameters, as described in Fig.2. In contrast, textual descriptions and infrared images convey coarse-grained information. Even in different tasks, they often follow similar feature extraction patterns, causing LoRA modules for text or infrared modalities to be similar across clients. When merging LoRA, existing FL [31, 35, 18] methods utilize the same aggregation strategy and ignores the difference within inter-modality complexity, thus failing to effectively integrate cross-modal abilities. Such disparities raise the key question: I) ***how to merge LoRA parameters under varying inter-modality complexity in cross-modal ReID tasks?***

In addition, another difficulty is the imbalance of intra-modality complexity. In cross-modal ReID, the complexity of a modality can vary across tasks due to differences in semantic gaps. For example, in the V–I task, both inputs are images, resulting in a relatively small semantic gap and lower visible LoRA complexity. However, V–T requires aligning visual features with textual semantics, which involves a much larger semantic gap and yields higher visible LoRA complexity. Such intra-modality complexity differences impact each task's contribution during modality-specific LoRA aggregation. Existing FL methods [19, 68, 16] assign aggregation weights by local data size or equally, assuming similar task difficulty. Relying solely on data size or equal weighting overlooks these differences, causing limited optimization for complex tasks and redundant parameters for simpler ones, degrading overall performance. This imbalance raises another key question: II) ***how to balance contributions across different cross-modal ReID tasks according to intra-modality complexity?***

Based on these challenges, we propose **MIRAGE** (**M**odality-heterogeneous **I**ntelligent federated **R**eID **A**ggregation with **G**lobal **E**fficiency) to integrate cross-modal knowledge from modality-

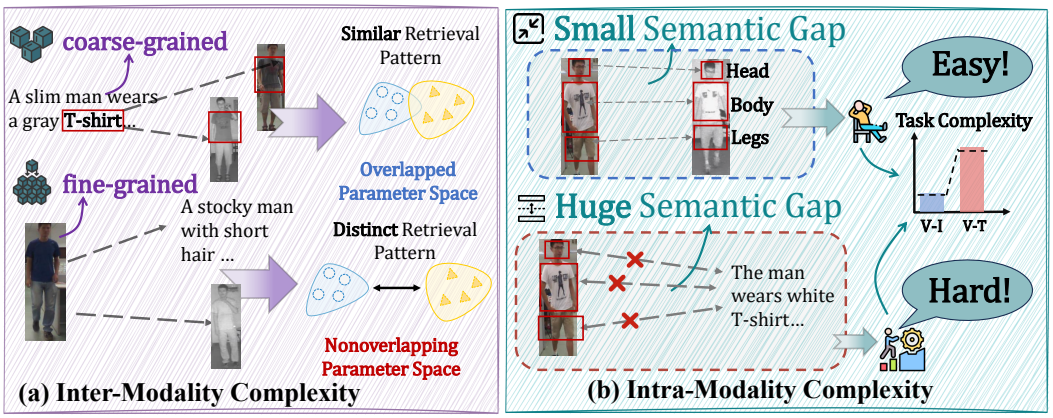


Figure 2: **inter- and intra-modality complexity.** (a) Different inter-modality complexity leads to different retrieval patterns and parameter space. (b) One modality in different cross-modal tasks faces various semantic gaps, resulting in different task complexity.

heterogeneous clients. To address question I), we design modality-specific LoRA aggregation strategies for clients with different modalities. Specifically, Merge with Selective Ranks (MSR) is used for visible LoRA, while Merge with Selectively Pruning (MSP) is applied to infrared and text LoRA. MSR selects important ranks according to the Frobenius norm to preserve crucial retrieval patterns, while reserving shared ranks for common knowledge. MSP prunes LoRA parameters to reduce intra-task conflicts. To address question II), we introduce Complexity-Aware Weighting (CAW) into adaptive aggregation, which quantifies task complexity and balances the contributions of different tasks. Our contributions are summarized as follows:

- **Task Contribution.** We introduce modality-heterogeneous federated ReID, a new task focusing on training unified cross-modal ReID framework which enables multiple cross-modal retrieval without needing for real multi-modal datasets and privacy leakage.
- **Methodological Contribution.** We propose a federated framework with clients holding different cross-modal data. Via dynamic merging and weighting strategies based on inter- and intra-modality complexity, our framework harmoniously integrate cross-modal abilities.
- **Empirical Contribution.** Extensive experiments on three cross-modal retrieval tasks demonstrate the superiority of our proposed MIRAGE.

2 Related Work

2.1 Person Re-identification

Person Re-identification (ReID) aims to match pedestrians across camera views [40, 36, 45, 70]. Early studies [1, 65, 76, 77, 24, 12] mainly focus on single-modal ReID, where query and gallery images share the same modality, typically RGB, advancing through deep feature learning [14, 48] and metric optimization [64, 25–27]. Cross-modal ReID [10, 11]further matches identities across modalities, such as RGB-infrared [60, 61, 63, 74] and RGB-text [20, 38, 21], enabling nighttime retrieval or description-based search. However, real-world queries may come from visible, infrared, or textual inputs, making existing single-modal or pairwise cross-modal methods insufficient. Recent unified ReID works [29, 4, 71, 50] address multi-modal inputs but rely on centralized training and costly multi-modal datasets, raising privacy and data collection concerns.

Federated ReID methods [53, 79, 80, 57] mitigate privacy risks by learning across distributed clients without sharing raw data. They also improve generalization across camera networks under domain shifts. However, most existing federated ReID methods are designed for visible images, limiting their applicability when queries involve infrared images or textual descriptions. To address this gap, our framework trains unified cross-modal ReID distributively, avoiding real multi-modal datasets while supporting robust retrieval across heterogeneous queries.

2.2 Low-Rank Adaptation

LoRA [17, 55, 3, 56] is a parameter-efficient fine-tuning method that represents model updates with two low-rank matrices inserted into existing layers, while keeping pretrained weights frozen. By optimizing only these low-rank matrices, LoRA greatly reduces trainable parameters and enables efficient adaptation with competitive performance [49, 9]. Recent studies [8, 34] show that LoRA parameters often contain redundancy that may contribute little or even harm generalization. Other works [33, 75, 62] further view each rank as an independent knowledge component, motivating selective preservation of key ranks for effective transfer. Despite its success in fine-tuning, LoRA remains underexplored in distributed training.

2.3 Federated Learning

Federated learning (FL) [37, 18, 35, 31, 28] enables collaborative model training across clients without sharing raw data. FedAvg [37] averages local model parameters and redistributes the global model for iterative training. Extensions such as FedProx [32] and SCAFFOLD [22] improve stability under heterogeneous data distributions. To reduce communication costs, recent works [51, 67, 2, 44] combine FL with LoRA. For example, FedIT [72] aggregates LoRA modules across clients, while FLoRA [51] concatenates client-specific LoRA components for noise-free aggregation. However, these methods usually assume shared tasks across clients, ignoring differences in objectives and task complexity. In modality-heterogeneous ReID, this causes incompatible client contributions. Our method addresses this by dynamically adapting aggregation strategies according to inter- and intra-modality complexity, enabling harmonious integration of diverse cross-modal knowledge.

3 Method

3.1 Preliminaries

Following the standard federated learning setup, we consider three clients, denoted as C_{V-I}, C_{V-T}, and C_{T-I}, corresponding to different cross-modal retrieval tasks. Each client owns a private dataset $D = \{x^i_{m_1}, y^i_{m_2}\}_{i=1}^N$ and two encoders f_{m_1}, f_{m_2}, where $m_1, m_2 \in \{vis, ir, txt\}$. Here, $x^i_{m_1}$ and $y^i_{m_2}$ denote the visible image, infrared image, or textual description of the i-th identity. To leverage existing cross-modal ReID methods and datasets, C_{V-I} and C_{V-T} adopt available cross-modal training algorithms, such as AGW [65] and IRRA [20]. Since T-I data and models are scarce, C_{T-I} uses a public V-T dataset, e.g., ICFG-PEDES [6], and generates pseudo infrared samples via channel augmentation [66]. A simple contrastive loss is then applied for coarse-grained T-I alignment.

LoRA is a low-rank adaptation method for efficiently fine-tuning large models. Given a pre-trained weight matrix $\mathbf{W} \in \mathbb{R}^{d\times k}$, LoRA introduces two trainable matrices $\mathbf{A} \in \mathbb{R}^{r\times k}$ and $\mathbf{B} \in \mathbb{R}^{d\times r}$, where $r \ll \min(d, k)$. The adapted weight is formulated as

$$\mathbf{W}' = \mathbf{W} + \mathbf{BA}, \tag{1}$$

where $\mathbf{W}$ is frozen and only $\mathbf{A}$ and $\mathbf{B}$ are optimized, greatly reducing trainable parameters while preserving adaptation capacity.

We build federated LoRA by inserting modality-specific LoRA modules into each client's encoders. During local training, clients optimize LoRA parameters on their private cross-modal data. During upload, only LoRA modules are transmitted to the server. The server aggregates LoRA modules belonging to the same modality across clients to obtain global LoRA parameters, which are then downloaded for the next training round. This enables effective knowledge sharing among heterogeneous modalities, and our work mainly focuses on the aggregation stage.

3.2 Complexity-Aware Weighting

In LoRA module, each rank corresponds to a row in the down-projection matrix A and a column in the up-projection matrix B. For example, $\mathbf{B} = [\mathbf{B}_1\ \mathbf{B}_2\ \cdots\ \mathbf{B}_r]$ and $\mathbf{A} = \left[\mathbf{A}_1^\top\ \mathbf{A}_2^\top\ \cdots\ \mathbf{A}_r^\top\right]^\top$. The computation of LoRA can be formulated as:

$$\mathbf{W}' = \mathbf{W} + \mathbf{BA} = \mathbf{W} + (\sum_{i=1}^{r} \mathbf{B}_i\mathbf{A}_i). \tag{2}$$

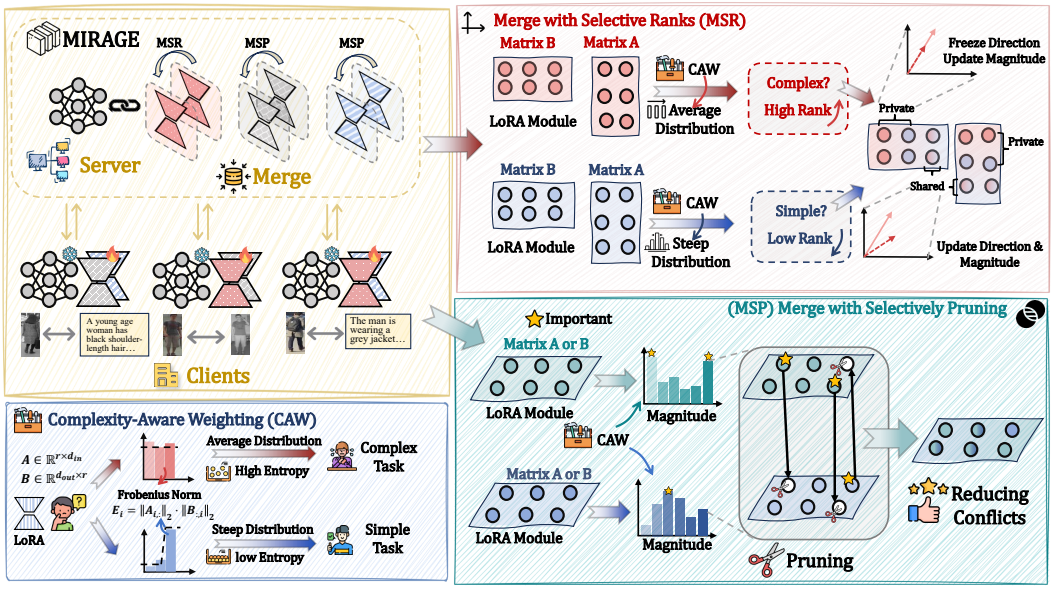


Figure 3: **The architecture of our proposed method.** MIRAGE consists of (a) Complexity-Aware Weighting(described in Sec. 3.2), (b) Merge with Selective Ranks(described in Sec. 3.3, (c) Merge with Selectively Pruning(described in Sec. 3.4.

Since each rank is computed independently, the parameters of each rank can be regarded as a functionally independent unit [75]. From a magnitude view, each rank produces a rank-1 update matrix, whose strength indicates its influence on the overall parameter update. For the i-th rank, this contribution is given by the outer product $\mathbf{B}_i\mathbf{A}_i$, and we measure its strength using the Frobenius norm: $E_i = \|\mathbf{B}_i\mathbf{A}_i\|_F$. A larger E_i indicates stronger influence along its parameter direction, suggesting higher importance.

Thus, the distribution of $\{E_i\}_{i=1}^{r}$ reflects how update capacity is allocated across ranks. If the values of E_i are **roughly equal**, many ranks contribute similarly, meaning the LoRA update involves multiple independent and important directions. Such updates are harder to approximate and therefore more **complex**. In contrast, a **steep** distribution concentrates most strength in a few ranks, where only a small subset of directions dominate and others have limited effect. This yields a **simpler** update that can be represented with fewer active ranks. Therefore, we measure update complexity by computing the Shannon entropy of the Frobenius norm distribution across ranks in each LoRA layer [33]. For the ℓ-th LoRA layer, we first normalize the energy of each rank and compute the layer wise entropy:

$$p_i^{(\ell)} = \frac{E_i^{(\ell)}}{\sum_{j=1}^{r} E_j^{(\ell)}}, \quad H^{(\ell)} = -\sum_{i=1}^{r} p_i^{(\ell)} \log p_i^{(\ell)}. \tag{3}$$

A high-entropy layer indicates that update strength is spread across ranks, suggesting more complex behavior, while a low-entropy layer concentrates influence in a few ranks and is structurally simpler. To assess task-level complexity, we average the entropy across all LoRA layers:

$$\bar{H} = \frac{1}{L}\sum_{\ell=1}^{L} H^{(\ell)}, \tag{4}$$

where L denotes the total number of LoRA layers in the client's model. This average entropy reflects the global difficulty of the local cross-modal retrieval task.

When aggregating LoRA modules from two clients with different tasks, we use $\bar{H}$ to estimate their relative task difficulties. Let $\bar{H}_1$ and $\bar{H}_2$ denote the average entropies of the two clients. We convert these difficulty scores into aggregation weights using a softmax function:

$$w_1 = \frac{exp(\bar{H}_1/\tau)}{exp(\bar{H}_1/\tau) + exp(\bar{H}_2/\tau)}, \quad w_2 = 1 - w_1, \tag{5}$$

where τ controls the sharpness of the weighting. This allows each client's LoRA module to contribute according to the intrinsic complexity of its task during aggregation.

3.3 Merge with Selective Ranks

Visible inputs carry fine-grained details, yet their LoRA update directions differ drastically across tasks. In V–T, updates align with textual semantics, while in V–I they focus on visual textures and infrared structures. These fundamentally different directions yield nearly disjoint LoRA subspaces, with principal ranks rarely overlapping.

Existing federated aggregation algorithms typically perform element-wise averaging of LoRA modules. However, because important-rank positions differ between tasks, naive averaging fails to preserve important cross-modal knowledge and dilutes the most critical patterns or ranks. To address this, our **Merge with Selective Ranks** (MSR) aggregation first identifies the **important ranks** for each task based on their Frobenius norm, and concatenates them to form a composite module:

$$\mathbf{A}_{concat}^{\top} = \left[\mathbf{A}_{imp}^{(V-T)^{\top}}, \mathbf{A}_{imp}^{(V-I)^{\top}}\right]^{\top}, \mathbf{B}_{concat} = \left[\mathbf{B}_{imp}^{(V-T)}, \mathbf{B}_{imp}^{(V-I)}\right]. \tag{6}$$

Nevertheless, due to the orthogonality between LoRA parameter subspaces, direct concatenation cannot transfer knowledge between V-I and V-T.

Therefore, MSR **reserves a shared rank subspace**. In practice, the aggregated LoRA matrices after MSR take the form:

$$\mathbf{A}_{MSR}^{\top} = \left[\mathbf{A}_{imp}^{(V-T)^{\top}}, \mathbf{A}_{shared}^{\top}, \mathbf{A}_{imp}^{(V-I)^{\top}}\right]^{\top}, \mathbf{B}_{MSR} = \left[\mathbf{B}_{imp}^{(V-T)}, \mathbf{B}_{shared}, \mathbf{B}_{imp}^{(V-I)}\right]. \tag{7}$$

The middle ranks $\mathbf{A}_{shared}$ and $\mathbf{B}_{shared}$ acts as the **knowledge bridge**, enabling MSR to preserve different task-specific update directions while enabling the construction of shared common knowledge.

For the non-shared part of the aggregation, MSR allocates ranks to each task according to the adaptive weights obtained from CAW, as described in Section 3.2. Let r_{total} denote the total ranks and r_{shared} the number of ranks reserved for shared knowledge. The allocation for each task in given by:

$$r_{V-T} = \lfloor (r_{\text{total}} - r_{\text{shared}}) \cdot w_1 \rfloor, \quad r_{V-I} = \lceil (r_{\text{total}} - r_{\text{shared}}) \cdot w_2 \rceil, \tag{8}$$

where w_1 and w_2 are the weights computed by CAW. For each task, the top-r_{V-T} and top-r_{V-I} important ranks are selected based on the Frobenius norm and directly inserted into the aggregated module. For the shared part, MSR selects r_{shared} ranks from the remaining important ranks of each task. In the shared subspace, the two tasks' parameters are averaged using the CAW weights:

$$\mathbf{A}_{shared} = w_1 \cdot \mathbf{A}_{rem}^{(V-T)} + w_2 \cdot \mathbf{A}_{rem}^{(V-I)}, \mathbf{B}_{shared} = w_1 \cdot \mathbf{B}_{rem}^{(V-T)} + w_2 \cdot \mathbf{B}_{rem}^{(V-I)}. \tag{9}$$

During subsequent federated training rounds, non-shared ranks only update their magnitude while freezing the directions to preserve important cross-modal retrieval abilities. As for a row $\mathbf{a}_i$ in matrix A in the non-shared part, the updating process can be described as:

$$\mathbf{a}_i^{(t+1)} = \mathbf{a}_i^{(t)} - \eta \Big[\big(\nabla_{\mathbf{a}_i} L \cdot \hat{\mathbf{a}}_i \big) \hat{\mathbf{a}}_i \Big], \tag{10}$$

where $\hat{\mathbf{a}}_i = \frac{\mathbf{a}_i}{\|\mathbf{a}_i\|_2}$ means the direction, t denotes the t-th epoch, and η is the learning rate. The same update rule is applied to the column $\mathbf{b}_i$ in matrix B corresponding to the same rank. Shared ranks can update both the magnitude and direction. Such design enables the shared subspace to continuously absorb and integrate knowledge from both V-T and V-I tasks, facilitating the construction of representations that are useful across modalities.

3.4 Merge with Selectively Pruning

Since infrared and text sources always contain coarse-grained information. In cross-modal pairs such as Text-Visible (T-V) and Text-infrared (T-I) as well as Infrared-Visible (I-V) and Infrared-Text (I-T), the learned LoRA updates tend to occupy overlapping parameter subspaces.

When aggregating text or infrared LoRA modules, the latent parameter spaces are not mutually irrelevant. Blindly concatenating their ranks leads to redundant parameter directions and hampers the model's overall performance. To address this, we propose the **Merge with Selectively Pruning** (MSP), which aggregates LoRA parameters in an element-size manner.

In LoRA modules, the absolute magnitude of each parameter can be taken as a measure of its importance. When merging two tasks with similar retrieval patterns, MSP reduces direct conflicts

between important parameters by introducing a dynamic pruning mechanism. The task complexity proportions w_1 and w_2, obtained from CAW as described in Section 3.2, are used to determine how many parameters in each task should be reserved, denoted as:

$$T_{imp} = N \cdot \frac{e^w - 1}{e^w + 1}, \tag{11}$$

where N is the total number of parameters in a LoRA layer and w is the task's complexity weight acquired via CAW. This saturating mapping adaptively scales the number of preserved parameters according to task difficulty, while preventing excessive dominance in highly complex scenarios. Based on T_{imp}, we define the *important parameter set* as

$$\mathcal{S} = \{v \mid v \text{ is among top-}T_{\text{imp}} \text{ absolute magnitudes}\}. \tag{12}$$

For task k and task q, we construct their mask based on the overlap with another task's important set:

$$M_k[v] = \begin{cases} 0, & v \in \mathcal{S}_q, \\ \frac{1}{2}, & v \notin \mathcal{S}_q, \end{cases} \tag{13}$$

and mask M_q can be acquired in the same way as M_k. Finally, the aggregated LoRA weights are obtained through an element-wise pruning fusion:

$$W_{agg} = M_k \odot W_k + M_q \odot W_q, \tag{14}$$

where k and q denote two clients conducting different tasks that would be merged and $\odot$ is the Hadamard product.

In subsequent local training for the client of task k, a Conflict-Aware Regularization (CAR) loss is employed to prevent the model from updating parameters located at crucial positions identified as important for another task q, which can be formally described as:

$$\mathcal{L}_{CAR} = \sum_{v \in \mathcal{S}_q} \|\Delta W[v]\|^2. \tag{15}$$

Therefore, the total objective becomes:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \mathcal{L}_{CAR}, \tag{16}$$

where $\mathcal{L}_{task}$ representing the specific cross-modal retrieval task. As for the client conducting task q, the total training loss can be obtained in the same way.

3.5 Limitation

Our LoRA aggregation relies on prior knowledge to capture parameter relationships within the same modality, which may hinder generalization when such priors are absent. In particular, in scenarios where modalities are weakly correlated, highly heterogeneous, or entirely unseen, the assumptions underlying MSR and MSP may not hold, potentially leading to suboptimal rank selection or conflicting pruning. This limitation may become more pronounced when extending MIRAGE to open-world settings, where new sensors, domains, or modality combinations can emerge without clear correspondence to existing modality relationships. Moreover, the predefined separation between fine-grained and coarse-grained modalities may not always be accurate, especially for tasks whose modality semantics vary across datasets or application scenarios. Future work could explore adaptive aggregation mechanisms that infer parameter relationships dynamically, extend the framework to handle unstructured or novel modalities, and reduce reliance on handcrafted or pre-specified priors.

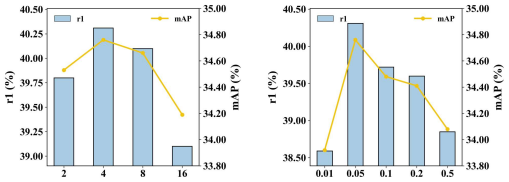
4 Experiments

4.1 Experimental Settings

Datasets and Evaluation Metrics. We validate visible-text and visible-infrared retrieval abilities of the proposed framework on two public person ReID datasets, *i.e.*, SYSU-MM01 [52] and RSTPReid [78]. In addition, we construct a Tri-SYSU-MM01 triple-modal dataset derived from SYSU-MM01 to further evaluate the text–infrared retrieval capability of MIRAGE. It is important to point out that client C_{T-I} is trained on ICFG-PEDES [6] with channel augmentation to generate pseudo infrared

Table 1: **Comparison with state-of-the-art FL methods** on SYSU-MM01, RSTPReid and Tri-SYSU-MM01, evaluating cross-modal retrieval capabilities (*i.e.*, visible-infrared, visible-text and infrared-text). Rank at r accuracy and mAP (%) are reported. The best and second-best performance are denoted in **bold** and underline.

Methods	SYSU-MM01			RSTPReid			Tri-SYSU-MM01			Average		
	r=1	r=5	mAP	r=1	r=5	mAP	r=1	r=5	mAP	r=1	r=5	mAP
FedAvg [37]	36.78	68.63	39.15	35.05	61.20	29.73	25.69	47.17	18.25	32.51	59.00	29.04
Traditional Federated Learning												
FedProx [32]	36.24	68.75	39.36	36.45	61.60	30.64	25.03	45.73	17.04	32.57	58.69	29.01
FedAvgM [16]	20.39	49.92	25.79	18.55	40.70	16.73	14.28	30.40	10.89	17.74	40.34	17.80
FedAdam [43]	24.64	54.31	27.19	36.90	61.75	29.09	28.85	**51.77**	16.78	30.13	55.94	24.35
FedYogi [43]	24.52	54.17	27.11	36.95	61.70	29.14	29.06	51.64	16.47	30.18	55.84	24.34
Scaffold [22]	37.45	71.57	40.14	40.75	66.95	32.86	28.79	48.65	19.51	35.66	62.39	30.84
Federated LoRA												
FFA-LoRA [44]	36.50	69.36	39.79	34.46	61.98	29.51	26.47	48.06	17.33	32.48	59.80	28.88
FLoRA [51]	42.36	74.58	44.34	30.89	54.96	26.90	26.55	47.26	19.02	33.27	58.93	30.09
FedSA-LoRA [13]	35.92	68.41	38.27	32.74	59.83	28.12	24.86	45.69	16.95	31.17	57.98	27.78
FRLoRA [58]	34.85	66.73	37.64	31.28	57.46	27.35	23.71	44.28	16.21	29.95	56.16	27.07
MIRAGE [Ours]	**44.45**	**75.70**	**46.00**	**46.20**	**71.50**	**37.37**	**30.27**	47.15	**20.92**	**40.31**	**64.78**	**34.76**



(a) Ablation of r_{shared} (b) Ablation of τ

Figure 4: **Hyperparameters study** of shared ranks in MSR r_{shared} in Eq. 8, and temperature τ in Eq. 5. The average Rank-1 accuracy (%) and mAP (%) are recorded.

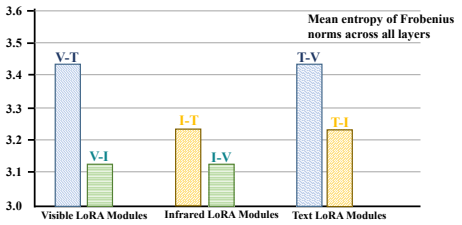


Figure 5: Mean entropy of modality-specific LoRA modules using Eq. 4. Higher entropy indicates greater complexity.

images, since there is almost no T-I datasets. Following existing works, we adopt the widely used Rank-k and mean Average Precision (mAP) as our evaluation metrics. The higher Rank-k and mAP indicates the higher performance. Detailed dataset descriptions are provided in the Appendix A.

Implementation Details. All image and text encoders are initialized from CLIP-ViT-B/16 and the CLIP text Transformer [42], respectively. We follow AGW [65] for C_{V-I} and IRRA [20] for C_{V-T}, and use channel augmentation [66] with a contrastive objective [39] for C_{T-I}. Only LoRA parameters are exchanged during federated training. More details are provided in Appendix B.

4.2 Comparison with State-of-the-Art Methods

We conduct extensive experiments comparing our framework with state-of-the-arts FL solutions on three different cross-modal retrieval tasks. As shown in Tab. 1, our method consistently outperforms the best existing approaches.

Comparison with traditional FL methods. Conventional federated learning (FL) [37, 32] assumes all clients perform identical tasks, focusing on mitigating client drift to stabilize the global model.

In our setting, clients provide modality-specific representations, and uniform aggregation is ill-suited for fusing heterogeneous cross-modal knowledge, thus limiting performance. In contrast, our dynamic merging mechanism (*i.e.*, MSR and MSP) effectively integrates diverse cross-modal retrieval abilities, achieving state-of-the-art results across all three retrieval tasks.

Comparison with federated LoRA methods.FFA-LoRA [44] updates only LoRA B matrices while freezing other parameters, greatly reducing trainable parameters but limiting capacity for complex tasks. FLoRA [51] concatenates LoRA ranks from clients to aggregate knowledge without loss, preserving all task-specific information but significantly increasing total ranks and trainable parameters. As reported in Tab. 1, FLoRA achieves outstanding performance on visible–infrared retrieval task, but does poorly in visible–text retrieval. This is because the direct rank-concatenation strategy ignores orthogonality between LoRA parameter spaces of tasks with distinct retrieval patterns,

hindering the transfer of shared knowledge. Moreover, under limited rank budgets, concatenating ranks from tasks with similar retrieval patterns introduces parameter redundancy and resulting in overfitting on simple task. Our method overcomes these issues, effectively aggregating diverse cross-modal knowledge and delivering strong performance across different retrieval scenarios.

4.3 Ablation Study

Ablation of Key Components. We evaluate the effectiveness of designed modules. As shown in Tab. 2, both proposed aggregation methods can boost overall performance. Specifically, using MSR and CAW leads to only a slight improvement, as the performance is still constrained by the naive aggregation of infrared and text LoRA modules. In contrast, employing MSP and CAW yields a substantial gain in overall cross-modal retrieval abilities. Notably, the combination of MSR and MSP without CAW suffers from uniform aggregation weights. This results in parameter redundancy on simple tasks and insufficient capacity for complex ones. Combining MSR, MSP, and CAW achieves the best results, indicating their complementary benefits in enhancing cross-modal knowledge integration.

Table 2: **Ablation study for key components**. Average Rank at r accuracy (%) and mAP (%) are reported.

MSR	MSP	CAW	Average		
			$r=1$	$r=5$	mAP
×	×	×	32.51	59.00	29.04
×	✓	✓	38.09	63.11	32.82
✓	×	✓	33.73	59.99	30.25
✓	✓	×	37.74	63.85	34.04
✓	✓	✓	**40.31**	**64.78**	**34.76**

Ablation of hyperparameters. We show the impact of hyperparameters r_{shared} (Eq. 8) and τ (Eq. 5) on performance in Fig. 4a and Fig. 4b. When merging orthogonal LoRA modules, r_{shared} balances client-specific and shared knowledge. A larger r_{shared} makes the method behave more like FedAvg, emphasizing global aggregation, whereas a smaller r_{shared} leads it closer to FLoRA. Our approach achieves a balance between these two extremes. τ controls the sharpness of complexity ratio across different tasks. When r_{shared}=4 and τ=0.05, our framework exhibits the best retrieval capabilities.

4.4 Further Discussion

We visualize the similarities across all ranks in LoRA modules belonging to the same modality but from different tasks. Take the visible LoRA modules shared between V-I and V-T as an example, we decompose the LoRA update $\Delta W = \mathbf{BA}$ into r individual rank components. The update corresponding to the z-th rank is computed as the outer product $\mathbf{U}_z = \mathbf{B}_{\cdot,z}\mathbf{A}_{z,\cdot}$. We calculate the similarity between the flattened update matrices of the i-th rank from V-T and the j-th rank from V-I as $S_{i,j} = \frac{\text{vec}(\mathbf{U}_i^{VT})\cdot\text{vec}(\mathbf{U}_j^{VI})}{\|\mathbf{U}_i^{VT}\|_F\|\mathbf{U}_j^{VI}\|_F}$. As shown in Fig. 6(a), visible LoRA ranks from the V-I and V-T clients exhibit low cross-task similarity, suggesting that they capture largely distinct update directions. Direct averaging may dilute these task-specific updates, motivating MSR to preserve important ranks from both clients. In contrast, Fig. 6(b) and Fig. 6(c) show higher cross-task similarity between infrared LoRA ranks from the I-V and I-T clients and between text LoRA ranks from the T-V and T-I clients, indicating overlapping update directions within each modality. These observations suggest that directly concatenating infrared and text LoRA ranks may introduce redundant directions, motivating MSP to selectively merge their parameters while reducing redundancy.

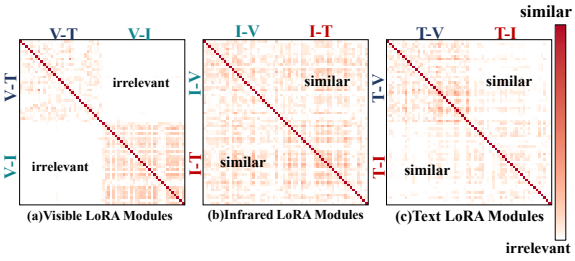


Figure 6: Visualization of similarities across all ranks in LoRA modules belonging to the same modality of different clients.

5 Conclusion

To the best of our knowledge, this is the first work to explore a unified cross-modal ReID framework supporting visible, infrared, and textual inputs. MIRAGE is achieved through aggregating modality-heterogeneous clients, without real multi-modal datasets or privacy risks. Our approach is built upon

three core modules: *Merge with Selective Ranks* (MSR), *Merge with Selectively Pruning* (MSP), and *Complexity-Aware Weighting* (CAW). MSR aggregates visible LoRA modules with orthogonal parameter space. And MSP efficiently fuses infrared and text LoRA parameters with similar retrieval patterns by avoiding potential conflicts. In both MSR and MSP, CAW decides aggregation weights of different cross-modal clients. Extensive experiments verify that our method can harmoniously integrate multiple cross-modal knowledge and achieve consistently strong performance.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (62501428, 62225113, 62361166629, 623B2080), the China Postdoctoral Science Foundation (2026T190427), the Hubei Provincial Natural Science Foundation of China (2025AFB219), National Research Foundation (NRF), Singapore, under Grant ID NRF-PA2025-NTU, and by Nanyang Technological University (NTU).

References

[1] Zechen Bai, Zhigang Wang, Jian Wang, Di Hu, and Errui Ding. Unsupervised multi-source domain adaptation for person re-identification. In *CVPR*, pages 12914–12923, 2021.

[2] Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. Lora-fair: Federated lora fine-tuning with aggregation and initialization refinement. In *ICCV*, pages 3737–3746, 2025.

[3] Massimo Bini, Leander Girrbach, and Zeynep Akata. Delora: Decoupling angles and strength in low-rank adaptation. *arXiv preprint arXiv:2503.18225*, 2025.

[4] Cuiqun Chen, Mang Ye, and Ding Jiang. Towards modality-agnostic person re-identification with descriptive query. In *CVPR*, pages 15128–15137, 2023.

[5] Haokun Chen, Yao Zhang, Denis Krompass, Jindong Gu, and Volker Tresp. Feddat: An approach for foundation model finetuning in multi-modal heterogeneous federated learning. In *AAAI*, volume 38, pages 11285–11293, 2024.

[6] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*, 2021.

[7] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, et al. Loramoe: Alleviating world knowledge forgetting in large language models via moe-style plugin. In *ACL*, pages 1932–1945, 2024.

[9] Chendi Ge, Xin Wang, Zeyang Zhang, Hong Chen, Jiapei Fan, Longtao Huang, Hui Xue, and Wenwu Zhu. Dynamic mixture of curriculum lora experts for continual multimodal instruction tuning. *arXiv preprint arXiv:2506.11672*, 2025.

[10] Yun-Peng Gong, Zhun Zhong, Zhiming Luo, Yan-Song Qu, Rongrong Ji, and Min Jiang. Cross-modality perturbation synergy attack for person re-identification. *ArXiv*, abs/2401.10090, 2024. URL `https://api.semanticscholar.org/CorpusID:267034769`.

[11] Yun-Peng Gong, Yongjie Hou, Jiang-Ming Shi, Kim Long Diep, and Min Jiang. A theory-inspired framework for few-shot cross-modal sketch person re-identification. In *AAAI Conference on Artificial Intelligence*, 2025. URL `https://api.semanticscholar.org/CorpusID:283245026`.

[12] Yunpeng Gong, Li-Qing Huang, and Lifei Chen. Person re-identification method based on color attack and joint defence. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4312–4321, 2021. URL `https://api.semanticscholar.org/CorpusID:249627387`.

[13] Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. Selective aggregation for low-rank adaptation in federated learning. *arXiv preprint arXiv:2410.01463*, 2024.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[15] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *ICCV*, pages 15013–15022, 2021.

[16] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1 (2):3, 2022.

[18] Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated learning. In *CVPR*, pages 10143–10153, 2022.

[19] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *CVPR*, pages 16312–16322. IEEE, 2023.

[20] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *CVPR*, pages 2787–2797, 2023.

[21] Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. Cross-modal cross-domain moment alignment network for person search. In *CVPR*, pages 10678–10686, 2020.

[22] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for on-device federated learning. *arXiv preprint arXiv:1910.06378*, 2(6), 2019.

[23] Zhi Kou, Jing Wang, Yu-Heng Jia, Biao Liu, and Xin Geng. Instance-dependent inaccurate label distribution learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36: 1425–1437, 2023. URL `https://api.semanticscholar.org/CorpusID:265156409`.

[24] Zhi Kou, Jing Wang, Jia-Wei Tang, Yu-Heng Jia, Bo-Yu Shi, and Xin Geng. Exploiting multi-label correlation in label distribution learning. *ArXiv*, abs/2308.01742, 2023. URL `https://api.semanticscholar.org/CorpusID:260438657`.

[25] Zhi Kou, Si Qin, Hailin Wang, Jing Wang, Ming-Kun Xie, Shuo Chen, Yuheng Jia, Tongliang Liu, Masashi Sugiyama, and Xin Geng. Label distribution learning with biased annotations assisted by multi-label learning. In *International Joint Conference on Artificial Intelligence*, 2025. URL `https://api.semanticscholar.org/CorpusID:281486411`.

[26] Zhi Kou, Yu-Cheng Xie, Hailin Wang, Junyang Chen, Jing Wang, Ming-Kun Xie, Shuo Chen, Yuheng Jia, Tong-Liang Liu, and Xin Geng. Rankmatch: A novel approach to semi-supervised label distribution learning leveraging rank correlation between labels. *Advances in Neural Information Processing Systems 38*, 2025. URL `https://api.semanticscholar.org/CorpusID:289033907`.

[27] Zhi Kou, Jun-Yang Chen, Xin-Qiang Cai, Xiaobo Xia, Ming-Kun Xie, Dong-Dong Wu, Biao Liu, Yuheng Jia, Xin Geng, Masashi Sugiyama, and Tat-Seng Chua. Positive-unlabeled reinforcement learning distillation for on-premise small models. *ArXiv*, abs/2601.20687, 2026. URL `https://api.semanticscholar.org/CorpusID:285101937`.

[28] Zhi Kou, Jun-Xian Wu, Wen-Ke Huang, Wen-Wen He, Ming-Kun Xie, Changwei Wang, Yuheng Jia, Di Jiang, Yang Liu, Xin Geng, and Qiang Yang. Fedharmony: Harmonizing heterogeneous label correlations in federated multi-label learning. *ArXiv*, abs/2604.28024, 2026. URL `https://api.semanticscholar.org/CorpusID:287915304`.

[29] He Li, Mang Ye, Ming Zhang, and Bo Du. All in one framework for multimodal re-identification in the wild. In *CVPR*, pages 17459–17469, 2024.

[30] Hongxia Li, Wei Huang, Jingya Wang, and Ye Shi. Global and local prompts cooperation via optimal transport for federated learning. In *CVPR*, pages 12151–12161, 2024.

[31] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *CVPR*, pages 10713–10722, 2021.

[32] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[33] Jian Liang, Wenke Huang, Xianda Guo, Guancheng Wan, Bo Du, and Mang Ye. Thanora: Task heterogeneity-aware multi-task low-rank adaptation. *arXiv preprint arXiv:2505.18640*, 2025.

[34] Jian Liang, Wenke Huang, Guancheng Wan, Qu Yang, and Mang Ye. Lorasculpt: Sculpting lora for harmonizing general and specialized knowledge in multimodal large language models. In *CVPR*, pages 26170–26180, 2025.

[35] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *NeurIPS*, 33:2351–2363, 2020.

[36] Jiawei Liu, Zhipeng Huang, Liang Li, Kecheng Zheng, and Zheng-Jun Zha. Debiased batch normalization via gaussian process for generalizable person re-identification. In *AAAI*, volume 36, pages 1729–1737, 2022.

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

[38] Ke Niu, Haiyang Yu, Mengyang Zhao, Teng Fu, Siyang Yi, Wei Lu, Bin Li, Xuelin Qian, and Xiangyang Xue. Chatreid: Open-ended interactive person retrieval via hierarchical progressive tuning for vision language models. *arXiv preprint arXiv:2502.19958*, 2025.

[39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[40] Jinjia Peng, Songyu Zhang, and Huibing Wang. Cde-learning: Camera deviation elimination learning for unsupervised person re-identification. In *AAAI*, volume 39, pages 6452–6460, 2025.

[41] Zhaopeng Peng, Xiaoliang Fan, Yufan Chen, Zheng Wang, Shirui Pan, Chenglu Wen, Ruisheng Zhang, and Cheng Wang. Fedpft: Federated proxy fine-tuning of foundation models. *arXiv preprint arXiv:2404.11536*, 2024.

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PmLR, 2021.

[43] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

[44] Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving lora in privacy-preserving federated learning. *arXiv preprint arXiv:2403.12313*, 2024.

[45] Yuan Sun, Yang Qin, Yongxiang Li, Dezhong Peng, Xi Peng, and Peng Hu. Robust multi-view clustering with noisy correspondence. *IEEE Transactions on Knowledge and Data Engineering*, 36:9150–9162, 2024. URL `https://api.semanticscholar.org/CorpusID:271131894`.

[46] Yuan Sun, Zhenwen Ren, Peng Hu, Dezhong Peng, and Xu Wang. Hierarchical consensus hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 26:824–836, 2024. URL `https://api.semanticscholar.org/CorpusID:258539654`.

[47] Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *CVPR*, pages 1522–1531, 2021.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[49] Lei Wang, Jieming Bian, Letian Zhang, and Jie Xu. Adaptive lora experts allocation and selection for federated fine-tuning. *arXiv preprint arXiv:2509.15087*, 2025.

[50] Zehan Wang, Yang Zhao, Haifeng Huang, Jiageng Liu, Aoxiong Yin, Li Tang, Linjun Li, Yongqi Wang, Ziang Zhang, and Zhou Zhao. Connecting multi-modal contrastive representations. *NeurIPS*, 36:22099–22114, 2023.

[51] Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. *NeurIPS*, 37:22513–22533, 2024.

[52] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, pages 5380–5389, 2017.

[53] Guile Wu and Shaogang Gong. Decentralised learning from independent multi-domain labels for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2898–2906, 2021.

[54] Jun-Xian Wu, Zhi Kou, Hongwei Zeng, Wen-Ke Huang, Biao Liu, Han-Lin Gu, Yuheng Jia, Di Jiang, Yang Liu, and Xin Geng. Trustworthy federated label distribution learning under annotation quality disparity. *ArXiv*, abs/2605.04827, 2026. URL `https://api.semanticscholar.org/CorpusID:287982052`.

[55] Yichen Wu, Hongming Piao, Long-Kai Huang, Renzhen Wang, Wanhua Li, Hanspeter Pfister, Deyu Meng, Kede Ma, and Ying Wei. Sd-lora: Scalable decoupled low-rank adaptation for class incremental learning. *arXiv preprint arXiv:2501.13198*, 2025.

[56] Jingwei Xu, Junyu Lai, and Yunpeng Huang. Meteora: Multiple-tasks embedded lora for large language models. *arXiv preprint arXiv:2405.13053*, 2024.

[57] Xin Xu, Chaoyue Ren, Wei Liu, Wenke Huang, Bin Yang, Zhixi Yu, and Kui Jiang. Positive style accumulation: A style screening and continuous utilization framework for federated dg-reid. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 8527–8536, 2025.

[58] Yunlu Yan, Chun-Mei Feng, Wangmeng Zuo, Rick Siow Mong Goh, Yong Liu, and Lei Zhu. Federated residual low-rank adaptation of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[59] Bin Yang, Mang Ye, Jun Chen, and Zesen Wu. Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In *ACMMM*, pages 2843–2851, 2022.

[60] Bin Yang, Jun Chen, and Mang Ye. Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In *ICCV*, pages 11069–11079, 2023.

[61] Bin Yang, Jun Chen, and Mang Ye. Shallow-deep collaborative learning for unsupervised visible-infrared person re-identification. In *CVPR*, pages 16870–16879, 2024.

[62] Yibo Yang, Xiaojie Li, Zhongzhu Zhou, Shuaiwen Song, Jianlong Wu, Liqiang Nie, and Bernard Ghanem. Corda: Context-oriented decomposition adaptation of large language models for task-aware parameter-efficient fine-tuning. *NeurIPS*, 37:71768–71791, 2024.

[63] Haoyu Yao, Bin Yang, Wenke Huang, Bo Du, and Mang Ye. Unsupervised visible-infrared person re-identification under unpaired settings. In *ICCV*, pages 11916–11926, 2025.

[64] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *CVPR*, pages 6210–6219, 2019.

[65] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *PAMI*, 44(6):2872–2893, 2021.

[66] Mang Ye, Zesen Wu, Cuiqun Chen, and Bo Du. Channel augmentation for visible-infrared re-identification. *PAMI*, 46(4):2299–2315, 2023.

[67] Liping Yi, Han Yu, Gang Wang, Xiaoguang Liu, and Xiaoxiao Li. pfedlora: Model-heterogeneous personalized federated learning with lora tuning. *arXiv preprint arXiv:2310.13283*, 2023.

[68] Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble. *arXiv preprint arXiv:2302.08888*, 2023.

[69] Xiaoyan Yu, Neng Dong, Liehuang Zhu, Hao Peng, and Dapeng Tao. Clip-driven semantic discovery network for visible-infrared person re-identification. *TMM*, 2025.

[70] Hong-Lin Yuan, Yuan Sun, Fei Zhou, Jing Wen, Shi-Hua Yuan, Xiao-Jian You, and Zhen-Wen Ren. Prototype matching learning for incomplete multi-view clustering. *IEEE Transactions on Image Processing*, 34:828–841, 2025. URL `https://api.semanticscholar.org/CorpusID:275776654`.

[71] Yajing Zhai, Yawen Zeng, Da Cao, and Shaofei Lu. Trireid: Towards multi-modal person re-identification via descriptive fusion model. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 63–71, 2022.

[72] Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. Towards building the federatedgpt: Federated instruction tuning. In *ICASSP*, pages 6915–6919. IEEE, 2024.

[73] Jie Zhang, Chen Chen, Bo Li, Lingjuan Lyu, Shuang Wu, Shouhong Ding, Chunhua Shen, and Chao Wu. Dense: Data-free one-shot federated learning. *NeurIPS*, 35:21414–21428, 2022.

[74] Yafei Zhang, Lingqi Kong, Huafeng Li, and Jie Wen. Weakly supervised visible-infrared person re-identification via heterogeneous expert collaborative consistency learning. In *ICCV*, pages 12659–12669, 2025.

[75] Ziyu Zhao, Tao Shen, Didi Zhu, Zexi Li, Jing Su, Xuwu Wang, Kun Kuang, and Fei Wu. Merging loras like playing lego: Pushing the modularity of lora to extremes through rank-wise clustering. *arXiv preprint arXiv:2409.16167*, 2024.

[76] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.

[77] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[78] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *ACMMM*, pages 209–217, 2021.

[79] Weiming Zhuang, Yonggang Wen, Xuesen Zhang, Xin Gan, Daiying Yin, Dongzhan Zhou, Shuai Zhang, and Shuai Yi. Performance optimization of federated person re-identification via benchmark analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 955–963, 2020.

[80] Weiming Zhuang, Yonggang Wen, and Shuai Zhang. Joint optimization in edge-cloud continuum for federated unsupervised person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 433–441, 2021.

A Dataset Details

SYSU-MM01 [52] is a widely used visible–infrared person re-identification dataset captured with 2 infrared and 4 visible cameras. It contains 395 identities for training, including 22,258 visible and 11,909 infrared images. In testing, query set has 96 identities with 3,803 infrared images, and gallery set consists of 301 randomly selected visible images.

RSTPReid [78] contains 20,505 images of 4,101 unique identities captured by 15 cameras. Each person has 5 distinct images from different cameras, and every image is paired with two textual descriptions. Following the official split, the dataset is divided into 3,701 identities for training, 200 for validation, and 200 for testing.

ICFG-PEDES [6] comprises 54,522 images belonging to 4,102 identities, with each image annotated by a textual description. The dataset is partitioned into a training set and a testing set. Specifically, the training partition contains 34,674 image-text pairs from 3,102 identities, while the remaining 1,000 identities, consisting of 19,848 pairs, are reserved for testing.

Tri-SYSU-MM01 is constructed based on the complete image set of SYSU-MM01 [52], and further supplements it with manually curated textual descriptions. It contains 29,003 visible images, 15,712 infrared images, and 44,715 text descriptions.The text annotation process for Tri-SYSU-MM01 consists of three steps: 1) we first manually created fine-grained visible-light textual descriptions for each pedestrian based on the visible images. 2) Then MLLM was used to perform text data augmentation in combination with images, enhancing the diversity of pedestrian descriptions. 3) Finally, for each image, we selected a corresponding textual description for each image and manually corrected noisy information in augmented text. Tri-SYSU-MM01 deliberately includes rich color-independent information, such as accessory locations and absent carried items, so that the annotations can better support diverse real-world retrieval scenarios.

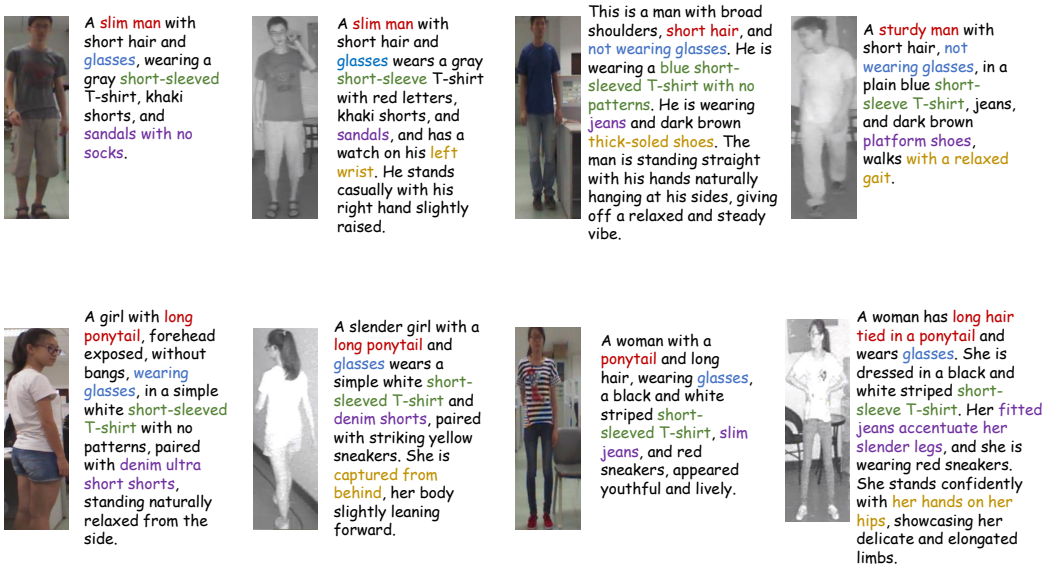


Figure 7: Visualization of the data samples from Tri-SYSU-MM01 Dataset.

B Implementation Details

In our framework, all pre-trained image encoders utilize CLIP-ViT-B/16 [42] and pre-trained text encoders employ CLIP text Transformer [42]. LoRA modules are injected into W_Q and W_V with the main parameters in model are frozen. All input images are resized to 384 × 128. The maximum length of the textual token sequence L is set to 77. In client C_{V-I} and C_{V-T}, they conduct AGW [65] and IRRA [20], where the loss function, optimizer and data augmentation are the same as mentioned in original papers. As for client C_{T-I}, a single contrastive loss [39] is used to achieve coarse-grained alignment and channel augmentation is adopted to generate pseudo infrared data [66]. We adopt the SGD optimizer with the initial learning rate of $3.5e-4$ to train this client. During training,

random horizontally flipping, random crop with padding, and random erasing are utilized for image augmentation. The entire training process is conducted over four communication rounds. In the first round, clients C_{V-I}, C_{V-T} and C_{T-I} are trained for 25, 40, and 10 epochs, respectively. In the subsequent rounds, these three clients are trained for 10, 10, and 3 epochs per round. The total ranks of LoRA modules is 32 and shared part in MSR is 4. The temperature in Equation 5 is set to 0.05. We perform our experiments on a single RTX4090 24GB GPU.

C Discussion

Concept Difference. Existing FL methods typically adopt uniform aggregation and assign contributions based on local data size or equally. In our modality-heterogeneous federated setting, such approaches overlook the difference of cross-modal retrieval tasks, hindering integration of diverse knowledge. Their weight assignment also cause overfitting on simpler tasks and neglecting of complex ones. MIRAGE addresses these issues via a dynamic aggregation strategy adapting to inter- and intra-modality complexity, enabling harmonious fusion of diverse cross-modal capabilities.

D Convergence Analysis

In this section, we provide a convergence analysis for the proposed MIRAGE aggregation. Since MIRAGE employs different aggregation strategies for different modalities, we explicitly decompose the aggregation error induced by MSR and MSP.

Let $\theta \in \mathbb{R}^d$ denote the collection of all trainable LoRA parameters. We decompose θ into two disjoint parts:

$$\theta = (\theta_{\text{vis}}, \theta_{\text{ct}}), \tag{17}$$

where θ_{vis} denotes the visible LoRA parameters aggregated by Merge with Selective Ranks (MSR), and θ_{ct} denotes the coarse-grained modality LoRA parameters, including infrared and text LoRA parameters, aggregated by Merge with Selectively Pruning (MSP). The global objective is defined as

$$F(\theta) = \sum_{k=1}^{N} p_k F_k(\theta), \tag{18}$$

where $F_k(\theta)$ is the local objective of client k, $p_k \geq 0$, and $\sum_{k=1}^{N} p_k = 1$.

At communication round t, each client performs local LoRA optimization on its private cross-modal data and uploads its modality-specific LoRA parameters to the server. The server applies MSR to visible LoRA modules and MSP to infrared and text LoRA modules. The overall MIRAGE update can be written as

$$\theta^{t+1} = \theta^t - \eta_t \sum_{k=1}^{N} w_k^t g_k^t + \xi_{\text{MSR}}^t + \xi_{\text{MSP}}^t + \xi_{\text{CAW}}^t, \tag{19}$$

where η_t is the learning rate, g_k^t is the stochastic gradient estimator of client k, and w_k^t is the aggregation weight generated by Complexity-Aware Weighting (CAW). The term ξ_{MSR}^t denotes the rank-selection approximation error introduced by MSR, ξ_{MSP}^t denotes the selective-pruning error introduced by MSP, and ξ_{CAW}^t denotes the discrepancy between the CAW-weighted aggregation direction and the ideal global aggregation direction.

For simplicity, we denote the total aggregation perturbation as

$$\xi_t = \xi_{\text{MSR}}^t + \xi_{\text{MSP}}^t + \xi_{\text{CAW}}^t. \tag{20}$$

Then Eq. (19) becomes

$$\theta^{t+1} = \theta^t - \eta_t \sum_{k=1}^{N} w_k^t g_k^t + \xi_t. \tag{21}$$

D.1 Preliminaries

We first make the following standard assumptions for federated stochastic optimization.

Assumption 1 (Smoothness). The global objective F is L-smooth, i.e.,

$$\|\nabla F(\theta) - \nabla F(\theta')\| \leq L\|\theta - \theta'\|, \quad \forall \theta, \theta' \in \mathbb{R}^d. \tag{22}$$

Assumption 2 (Unbiased stochastic gradients). For each client k, the stochastic gradient estimator satisfies

$$\mathbb{E}[g_k^t \mid \theta^t] = \nabla F_k(\theta^t). \tag{23}$$

Assumption 3 (Bounded variance). There exists a constant $\sigma^2 > 0$ such that

$$\mathbb{E}\left[\|g_k^t - \nabla F_k(\theta^t)\|^2 \mid \theta^t\right] \leq \sigma^2, \quad \forall k, t. \tag{24}$$

Assumption 4 (Bounded stochastic gradients). There exists a constant $G > 0$ such that

$$\mathbb{E}\|g_k^t\|^2 \leq G^2, \quad \forall k, t. \tag{25}$$

D.2 Bounded Error of MSR

MSR is applied to visible LoRA modules. For a LoRA layer, the update matrix can be decomposed into rank-wise components:

$$\Delta W = BA = \sum_{i=1}^{r} B_i A_i, \tag{26}$$

where $B_i A_i$ denotes the rank-i update component. MSR selects important ranks according to the Frobenius norm $\|B_i A_i\|_F$ and reserves a small number of shared ranks for cross-task knowledge transfer.

Let $\mathcal{R}_{\text{sel}}^{(\ell),t}$ denote the set of selected private and shared ranks in the ℓ-th LoRA layer at round t. The approximation error of MSR is caused by the unselected ranks. Therefore, it can be bounded by the remaining rank energy:

$$\|\xi_{\text{MSR}}^t\|^2 \leq C_{\text{MSR}} \sum_{\ell=1}^{L} \sum_{i \notin \mathcal{R}_{\text{sel}}^{(\ell),t}} \left\| B_i^{(\ell),t} A_i^{(\ell),t} \right\|_F^2, \tag{27}$$

where L is the number of LoRA layers and $C_{\text{MSR}} > 0$ is a constant. Since MSR selects ranks with the largest Frobenius norm and the LoRA parameters are finite-dimensional and bounded during training, there exists a constant $\epsilon_{\text{MSR}} > 0$ such that

$$\mathbb{E}\|\xi_{\text{MSR}}^t\|^2 \leq \epsilon_{\text{MSR}}^2, \quad \forall t. \tag{28}$$

D.3 Bounded Error of MSP

MSP is applied to infrared and text LoRA modules. These coarse-grained modality LoRA modules tend to occupy more similar parameter subspaces across different cross-modal tasks. Therefore, MSP performs element-wise selective pruning to reduce conflicts between important parameters.

Let W_k^t and W_q^t denote two LoRA parameter tensors to be merged by MSP, and let M_k^t and M_q^t be the pruning masks generated according to the important parameter sets of the other task. The MSP aggregation is given by

$$W_{\text{MSP}}^t = M_k^t \odot W_k^t + M_q^t \odot W_q^t, \tag{29}$$

where $\odot$ denotes element-wise multiplication.

The pruning error is caused by the masked-out parameters. Thus, it can be bounded as

$$\|\xi_{\text{MSP}}^t\|^2 \leq C_{\text{MSP}} \left(\|(1 - M_k^t) \odot W_k^t\|^2 + \|(1 - M_q^t) \odot W_q^t\|^2 \right), \tag{30}$$

where $C_{\text{MSP}} > 0$ is a constant. Since MSP prunes parameters according to their importance and the LoRA parameters are bounded during training, there exists a constant $\epsilon_{\text{MSP}} > 0$ such that

$$\mathbb{E}\|\xi_{\text{MSP}}^t\|^2 \leq \epsilon_{\text{MSP}}^2, \quad \forall t. \tag{31}$$

D.4 Bounded Discrepancy of CAW

CAW computes aggregation weights from the entropy of rank-wise Frobenius norm distributions. Since the weights are obtained by a softmax function, they satisfy

$$w_k^t \geq 0, \quad \sum_{k=1}^{N} w_k^t = 1. \tag{32}$$

Therefore, the CAW-weighted update direction is a convex combination of local update directions.

Let p_k denote the ideal global objective weight and w_k^t denote the CAW-generated weight. The discrepancy between the CAW-weighted update and the ideal global update can be written as

$$\xi_{\mathrm{CAW}}^t = \eta_t \sum_{k=1}^{N} (p_k - w_k^t) g_k^t. \tag{33}$$

By Assumption 4, we have

$$\mathbb{E}\|\xi_{\mathrm{CAW}}^t\|^2 \leq \eta_t^2 G^2 \left(\sum_{k=1}^{N} |p_k - w_k^t| \right)^2. \tag{34}$$

Since both p_k and w_k^t are normalized weights, the above quantity is bounded. Therefore, there exists a constant $\epsilon_{\mathrm{CAW}} > 0$ such that

$$\mathbb{E}\|\xi_{\mathrm{CAW}}^t\|^2 \leq \epsilon_{\mathrm{CAW}}^2, \quad \forall t. \tag{35}$$

Lemma 1 (Bounded MIRAGE aggregation perturbation). The total perturbation introduced by MIRAGE aggregation is bounded:

$$\mathbb{E}\|\xi_t\|^2 \leq \epsilon_{\mathrm{agg}}^2, \tag{36}$$

where

$$\epsilon_{\mathrm{agg}}^2 = 3\left(\epsilon_{\mathrm{MSR}}^2 + \epsilon_{\mathrm{MSP}}^2 + \epsilon_{\mathrm{CAW}}^2\right). \tag{37}$$

Proof. Using Eq. (20) and the inequality $\|a+b+c\|^2 \leq 3(\|a\|^2 + \|b\|^2 + \|c\|^2)$, we have

$$\begin{aligned} \mathbb{E}\|\xi_t\|^2 &= \mathbb{E}\left\|\xi_{\mathrm{MSR}}^t + \xi_{\mathrm{MSP}}^t + \xi_{\mathrm{CAW}}^t\right\|^2 \\ &\leq 3\mathbb{E}\|\xi_{\mathrm{MSR}}^t\|^2 + 3\mathbb{E}\|\xi_{\mathrm{MSP}}^t\|^2 + 3\mathbb{E}\|\xi_{\mathrm{CAW}}^t\|^2 \\ &\leq 3\left(\epsilon_{\mathrm{MSR}}^2 + \epsilon_{\mathrm{MSP}}^2 + \epsilon_{\mathrm{CAW}}^2\right). \end{aligned} \tag{38}$$

This completes the proof. □

D.5 Convergence of MIRAGE

Theorem 1 (Convergence of MIRAGE). Under Assumptions 1–4 and Lemma 1, if the learning rate satisfies $\eta_t = \eta = O(1/\sqrt{T})$, then MIRAGE satisfies

$$\min_{0 \leq t < T} \mathbb{E}\|\nabla F(\theta^t)\|^2 \leq O\left(\frac{1}{\sqrt{T}}\right) + O\left(\epsilon_{\mathrm{agg}}^2\right). \tag{39}$$

Therefore, MIRAGE converges to a first-order stationary point up to a bounded aggregation error.

Proof. By the L-smoothness of F, we have

$$F(\theta^{t+1}) \leq F(\theta^t) + \left\langle \nabla F(\theta^t), \theta^{t+1} - \theta^t \right\rangle + \frac{L}{2}\|\theta^{t+1} - \theta^t\|^2. \tag{40}$$

Using Eq. (21), we have

$$\theta^{t+1} - \theta^t = -\eta_t \sum_{k=1}^{N} w_k^t g_k^t + \xi_t. \tag{41}$$

Define

$$G^t = \sum_{k=1}^{N} w_k^t g_k^t. \tag{42}$$

Then Eq. (41) becomes

$$\theta^{t+1} - \theta^t = -\eta_t G^t + \xi_t. \tag{43}$$

Substituting Eq. (43) into Eq. (40), we obtain

$$\begin{aligned} F(\theta^{t+1}) \leq F(\theta^t) - \eta_t \left\langle \nabla F(\theta^t), G^t \right\rangle + \left\langle \nabla F(\theta^t), \xi_t \right\rangle \\ + \frac{L}{2} \left\| -\eta_t G^t + \xi_t \right\|^2. \end{aligned} \tag{44}$$

Taking expectation and using the unbiasedness and bounded variance assumptions, we obtain

$$\mathbb{E}\|G^t\|^2 \leq C_g \mathbb{E}\|\nabla F(\theta^t)\|^2 + C_\sigma \sigma^2, \tag{45}$$

where C_g and C_σ are positive constants depending on the aggregation weights. Since CAW produces normalized non-negative weights, G^t is a convex combination of local stochastic gradients. Therefore, under the standard descent-direction condition, there exists a constant $c > 0$ such that

$$\mathbb{E}\left[\left\langle \nabla F(\theta^t), G^t \right\rangle\right] \geq c \mathbb{E}\|\nabla F(\theta^t)\|^2. \tag{46}$$

Moreover, by Young's inequality and Lemma 1,

$$\mathbb{E}\left[\left\langle \nabla F(\theta^t), \xi_t \right\rangle\right] \leq \frac{c\eta_t}{2} \mathbb{E}\|\nabla F(\theta^t)\|^2 + \frac{1}{2c\eta_t} \epsilon_{\text{agg}}^2. \tag{47}$$

Using $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, Eq. (45), and Lemma 1, we have

$$\mathbb{E}\left\| -\eta_t G^t + \xi_t \right\|^2 \leq 2\eta_t^2 \left(C_g \mathbb{E}\|\nabla F(\theta^t)\|^2 + C_\sigma \sigma^2 \right) + 2\epsilon_{\text{agg}}^2. \tag{48}$$

Combining Eqs. (44)–(48), and choosing a sufficiently small learning rate, we obtain

$$\mathbb{E}[F(\theta^{t+1})] \leq \mathbb{E}[F(\theta^t)] - C_1 \eta_t \mathbb{E}\|\nabla F(\theta^t)\|^2 + C_2 \eta_t^2 \sigma^2 + C_3 \epsilon_{\text{agg}}^2, \tag{49}$$

where C_1, C_2, and C_3 are positive constants.

Summing Eq. (49) from $t = 0$ to $T - 1$ and using the lower bound $F^\star$ of F, we get

$$C_1 \sum_{t=0}^{T-1} \eta_t \mathbb{E}\|\nabla F(\theta^t)\|^2 \leq F(\theta^0) - F^\star + C_2 \sigma^2 \sum_{t=0}^{T-1} \eta_t^2 + C_3 T \epsilon_{\text{agg}}^2. \tag{50}$$

Therefore,

$$\min_{0 \leq t < T} \mathbb{E}\|\nabla F(\theta^t)\|^2 \leq O\left(\frac{F(\theta^0) - F^\star}{\sum_{t=0}^{T-1} \eta_t} \right) + O\left(\frac{\sigma^2 \sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t} \right) + O\left(\frac{T \epsilon_{\text{agg}}^2}{\sum_{t=0}^{T-1} \eta_t} \right). \tag{51}$$

With $\eta_t = \eta = O(1/\sqrt{T})$, we have $\sum_{t=0}^{T-1} \eta_t = O(\sqrt{T})$ and $\sum_{t=0}^{T-1} \eta_t^2 = O(1)$. Absorbing constants into the perturbation term, we obtain

$$\min_{0 \leq t < T} \mathbb{E}\|\nabla F(\theta^t)\|^2 = O\left(\frac{1}{\sqrt{T}} \right) + O\left(\epsilon_{\text{agg}}^2 \right). \tag{52}$$

This completes the proof. □

Corollary 1 (Exact convergence under vanishing aggregation error). If the aggregation perturbation satisfies $\epsilon_{\text{agg}}^2 = O(1/\sqrt{T})$, then MIRAGE preserves the standard convergence rate of federated stochastic optimization:

$$\min_{0 \leq t < T} \mathbb{E}\|\nabla F(\theta^t)\|^2 = O\left(\frac{1}{\sqrt{T}} \right). \tag{53}$$

The above analysis shows that MIRAGE does not break the convergence behavior of federated LoRA optimization. Since MSR and MSP are applied to different modality-specific LoRA modules, their aggregation errors are analyzed separately and then combined into a unified bounded perturbation term. The effect of CAW is also bounded because it produces normalized non-negative aggregation weights. Therefore, MIRAGE converges to a first-order stationary point up to a bounded aggregation error. When the rank-selection error in MSR, the pruning error in MSP, and the weighting discrepancy in CAW are sufficiently small or decrease during training, MIRAGE achieves the standard $O(1/\sqrt{T})$ convergence rate for smooth non-convex objectives.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main claims of our work, including an overview of MIRAGE. The main contribution and limitation can be found in Sec. 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We clearly discuss the limitations of MIRAGE's reliance on prior modality relationships in Sec. 3.5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper provides the assumptions and proof for the convergence analysis of the proposed method in the Appendix D. The theoretical result is limited to convergence behavior and does not claim to fully explain the empirical effectiveness of the method.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides the key experimental details needed to reproduce the main results, including datasets, model settings, training protocols, hyperparameters, evaluation metrics, and implementation details in Sec. 4 and Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: The source code and data will be publicly released upon publication.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify the key training and testing details needed to understand the results, including the datasets used to simulate modality-heterogeneous scenarios, data split settings, training configurations, hyperparameters, optimizer choices, and evaluation protocols. All details can be found in Sec. 4 and Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars or formal statistical significance tests. However, to obtain stable performance estimates, we evaluate MIRAGE 10 times on each cross-modal dataset and report the average performance, following the protocol used in related works.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to Appendix B for details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conforms to the NeurIPS Code of Ethics. We use standard research datasets and do not involve human subjects, private or sensitive personal data, or applications that raise significant ethical concerns.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work does not release data or models with a high risk of misuse, such as pre-trained language models, image generators, or scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly credit the creators and original owners of the datasets, models, and code used in our work through citations and references, and we respect their licenses and terms of use. The details of datasets and models can be seen in Appendix A and Appendix B.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes. We will provide documentation alongside the released code and data, including usage instructions, dataset descriptions, preprocessing details, and reproduction scripts.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: None of the core methods, including any important, original, or non-standard components, rely on LLMs. We also do not use LLMs to generate data, experimental results, or other scientific content that affects the methodology or conclusions.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.