Distributionally Robust Alignment for Medical Federated Vision-Language Pre-training Under Data Heterogeneity

Anonymous authors
Paper under double-blind review

Abstract

Vision-language pre-training (VLP) has emerged as an effective scheme for multimodal representation learning, but its reliance on large-scale multimodal data poses significant challenges for medical applications. Federated learning (FL) offers a promising solution to scale up the dataset for medical VLP while preserving data privacy. However, we observe that client data heterogeneity in real-world scenarios could cause models to learn biased cross-modal alignment during local pre-training. This would limit the transferability of the federally learned representation model on downstream tasks. To address this challenge, we propose **Fed**erated **D**istributionally **R**obust **A**lignment (FedDRA), a framework for federated VLP that achieves robust vision-language alignment under heterogeneous conditions. Based on client datasets, we construct a distribution family that encompasses potential test-time domains, and apply a distributionally robust framework to optimize the pre-trained model's performance across this distribution space. This approach bridges the gap between pre-training samples and downstream applications. To avoid over-fitting on client-specific information, we use anchor representation from the global model to guide the local training, and adopt a two-stage approach to first tune deeper layers before updating the entire network. Extensive experiments on real-world datasets demonstrate FedDRA's effectiveness in enhancing medical federated VLP under data heterogeneity. Our method also adapts well to various medical pre-training methods.

1 Introduction

Vision-language pre-training (VLP) learns transferable multimodal representations by extracting latent semantics from large-scale image-text pairs, where the dataset scale largely impacts the performance of the learned model (Oquab et al., 2023). However, scaling up multimodal pre-training datasets is a non-trivial challenge especially for medical applications, due to privacy concerns and regulations of patient data sharing (Ladbury et al., 2023). Recent work has explored federated learning as a solution to leverage data across multiple medical institutions while preserving privacy (Lu et al., 2023).

However, in real-world scenarios, datasets collected from different institutes are always heterogeneous. For example, hospitals in tropical regions receive a high proportion of pneumonia patients, whereas those in colder climates may see more pneumothorax cases (Mendogni et al., 2020). This data heterogeneity is not only a long-standing problem in classical federated learning (Ghosh et al., 2019; Huang et al., 2022), but a practical challenge that impedes the deployment of medical VLP in the federated learning setting. Current medical VLP methods often focus on learning a modality-shared latent space, where their training multimodal data pairs are well-aligned. However, such learned cross-modal alignments may not be transferable to data from unseen distributions. As shown in Fig. 1, this will harm the performance of the federally pre-trained model, which is aggregated from client models trained on heterogeneous local datasets.

We start by investigating how data heterogeneity affects the performance of federally pre-trained VL models. In classic medical VLP (Wang et al., 2022; Bannur et al., 2023), the model learns cross-modal alignment through maximizing the mutual information of the two modalities on its observed training data. In federated settings, as shown in Fig. 1, this approach often learns local models that overfit client-specific information.

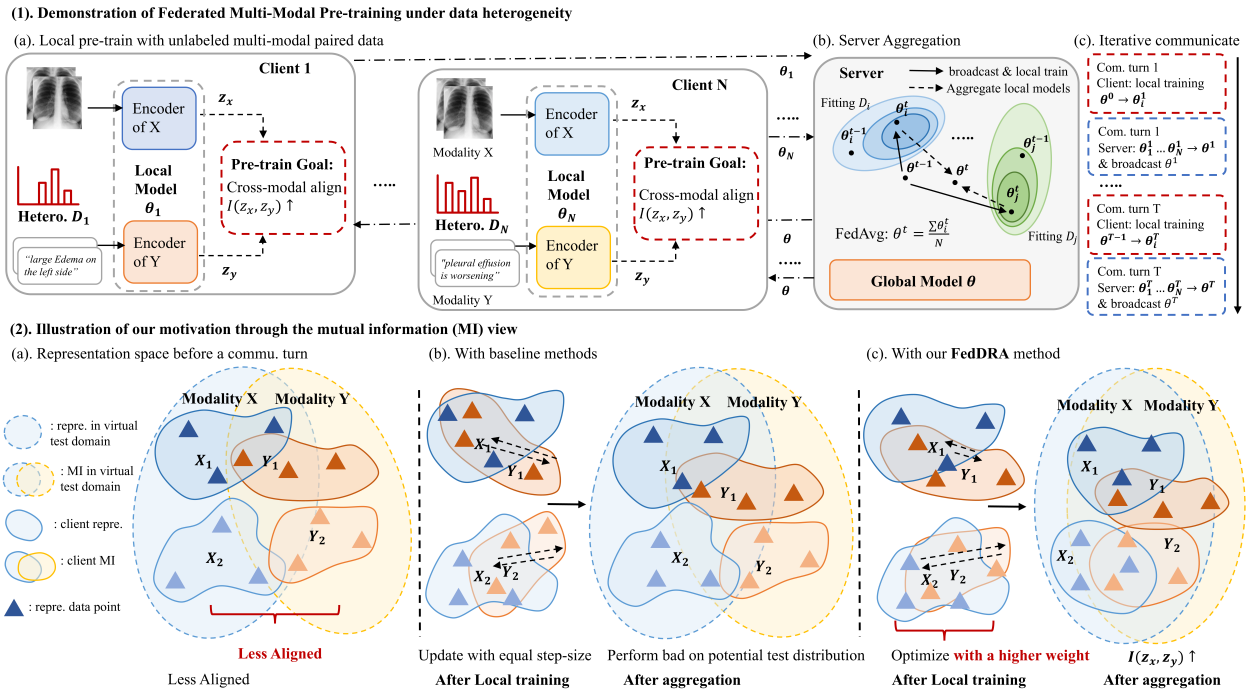

Figure 1: (1) We aim to tackle the data-hungry problem in VLP via federally utilizing private multi-modal paired data. Pre-training local models on heterogeneous client datasets may overfit the observed local data. (2). The naive method ignores the disparity between local distribution and potential testing distribution, and directly averaging local pre-trained models would obtain a less generalizable model. In contrast, our method optimizes model performance based *on a family of potential testing distributions*, dynamically given weights to schedule the local training, and obtain a more transferable model.

Also, averaging these local models' parameters cannot always produce a model with generalizable cross-modal alignment. Secondly, biased deep layers, which overfit multi-modal correlations of local datasets, would prevent the model from learning transferable and diverse semantics during local training.

In this paper, we propose a **Fed**erated **D**istributionally **R**obust **A**lignment (FedDRA) framework, to learn transferable cross-modal alignment under data heterogeneity. Our key idea is to maximize cross-modal mutual information with distributional robustness. Specifically, to bridge the gap between the downstream testing domain and pre-training samples, we construct a set of distributions based on client distributions. We then employ a decentralized distributionally robust optimization method to iteratively improve the pre-trained model's performance on this set. To alleviate the negative effect of over-fitting client-specific information, we maintain a global model to provide anchor representations for guiding local training and utilize a two-stage training schema to tune deep layers before updating the whole network.

Our contributions primarily focus on:

- We for the first time tackle the problem of medical VLP under the federated setting by utilizing heterogeneous multi-modal datasets from different institutes. We conduct empirical studies to analyze the influence of data heterogeneity on federated multi-modal learning.

- We propose FedDRA to address the data heterogeneity challenge in federated VLP to obtain transferable cross-modal alignment. It iteratively optimizes model performance on a distribution family and uses a two-stage global-guided local training strategy to reduce over-fitting on client-specific patterns.

Table 1: Comparison between related works and our proposed FedDRA for federated vision-language pre-training. We organize the related works based on task settings and technical points:(1) Heterogeneous client datasets. (2) Multi-modal paired datasets. (3) Server-computation-free. (4) No communication on training data.(5) Introduce global constraints. (6) Distributionally robust framework.

Related Work	**Task Settings**				**Related Technical Points**	
	Hetero.	Multi-Modal	Server Comp.-Free	Feat. Commu.-Free	Global Const.	Dist. Robust
Zhuang et al. (2021)	✓		✓	✓	✓	
Zhuang et al. (2022)	✓		✓	✓		✓
Zhang et al. (2023)	✓		✓		✓	
Yan et al. (2023)	✓			✓		
Lu et al. (2023)		✓	✓	✓		
Ours	✓	✓	✓	✓	✓	✓

- Experiment results show the effectiveness of our method in learning multi-modal representations under the federated setting for various downstream tasks, including image-text retrieval, classification, and segmentation.

2 Related Work

Medical Vision-Language Pre-training. Pre-training multi-modal models on large-scale datasets and then transferring learned knowledge to downstream tasks has become a popular approach to leverage diverse semantics contained in multi-modal unlabeled data (Li et al., 2022b; Bao et al., 2022; Radford et al., 2021). Current works aim to learn a shared latent space to connect the representations of each modality, leveraging a wide range of self-supervised learning methods, i.e., contrastive learning (Radford et al., 2021; Chen et al., 2020) and multi-modal fusion (Li et al., 2021a; Chen et al., 2022). Medical multi-modal pre-training tasks are often conducted on vision-based datasets, especially vision-language pre-training. (Zhang et al., 2022) first utilizes an image-text contrastive loss to align visual and language representations. (Huang et al., 2021) aligns fine-grained cross-modal representations (Huang et al., 2021) through a word-patch contrastive loss, and has improved the performance on fine-grained vision tasks. Furthermore, recent work (Wang et al., 2022; Bannur et al., 2023) incorporates medical domain knowledge to mitigate the misalignment during pre-training. However, almost all of the current methods still rely on large-scale pre-training datasets, which impede their adaptaion to modalities with limited training samples and deployment in real-world scenarios.

Heterogeneity in Self-Supervised Federated Learning. Federated self-supervised learning aims to leverage diverse semantics in local unlabeled datasets in a decentralized and privacy-preserved way. One of the key challenges of federated learning is data heterogeneity (Li & Wang, 2019; Collins et al., 2021; Li et al., 2021b), and has been long discussed in the uni-modality scenarios. Typically, (Zhang et al., 2023; Huang et al., 2022; Li et al., 2022a) employ additional communications on local data representations to increase sample diversity. Such methods fail to protect data privacy, which is a vital concern in medical applications. On the other hand, (Zhuang et al., 2021; Li et al., 2021b) utilizes server model to constrain the update of local models, (Yan et al., 2023) utilizes the mask-autoencoder to handle heterogeneity, (Zhuang et al., 2022; Li et al., 2021b; Han et al., 2022) considers distillation-based methods yet ignores the direct modeling of cross-modal alignment. However, these uni-modal self-supervised learning methods have not accounted for the modality gap (Zhang et al., 2024b) between multi-modal data. While uni-modal self-supervised learning aims to learn robust features (Radford et al., 2021), multi-modal learning need also align input modalities to maximize the mutual information between their representations (Su et al., 2023). Recent advances Lu et al. (2023) have verified that federated learning can be utilized to scale up the pre-training dataset. However, this work hasn't considered the harm of data heterogeneity issue (Ghosh et al., 2019). While the learned local models can be biased and over-reliance on spurious correlations (Saab et al., 2022) that are client-specific, distributionally robust optimization (Deng et al., 2020) framework can alleviate these issues by optimizing the group-wise worst-case performance on given objective (Liu et al., 2022), thus this idea can be flexibly adapted to various of federated learning tasks Han et al. (2023); Rehman et al. (2023); Capitani et al. (2024).

3 Problem Formulation

Formulation of Pre-training Dataset and Heterogeneity. In this paper, we consider the multi-modal datasets with two modalities X, Y, e.g., image and text modalities. Following (Su et al., 2023), we assume sample x of modality X and sample y of modality Y are generated from latent semantics through implicit mappings that are consistent across all clients. For instance, disease labels of a given X-ray image and radiology-report pair, are latent semantic variables that connect these two modalities. That's because these labels determine the pathology region of the radiology image and corresponding description in the diagnosis report. In federated learning, we consider N clients, each has a own local dataset, forming a group $\mathcal{C}$. We assume that each client has a correspond data distribution $D_i, i \in [1, \ldots, N]$, and data samples are given as $(x, y) \sim D_i$. In real-world scenarios, the distributions D_i of local datasets vary across clients, introducing data heterogeneity that can negatively impact federated learning performance.

To obtain a generalizable model that performs well on testing domain, we often consider a virtual global dataset with data distribution $D_{\mathcal{T}}$ (Zhang et al., 2024a). In real-world setting, testing domains are often out-of-distribution (OOD), not limited to the pre-training local datasets. For example, a medical multi-modal model might be pre-trained on data from routine clinical practice and then transferred to tasks utilizing datasets collected during the COVID-19 pandemic. Therefore, we consider a family of global data domains that includes distribution shifts, which can be written in a form of uncertainty set: $Q^c : \{Q : D_f(Q \| D^{\mathcal{C}}) \leq \rho\}$, where $D^{\mathcal{C}}$ is the distribution when entire data is grouped based on group $\mathcal{C}$, D_f is the f-divergence of two distributions. $\rho \in R^+$ is the uncertainty radius, a larger ρ introduces more unseen distributions.

Federated Vision-Language Pre-training. Given N clients and their local datasets, federated learning aims to utilize the client dataset to train a generalizable model in a privacy-preserved way. It iteratively trains local models on the client side and aggregates (e.g., FedAvg strategy simply averages model parameters) them on the server. For each communication turn r, each client learns a local model through E update steps, and sends it to the server, and overwrites the local model with the aggregated model sent back from the server. Specifically, Federated Multi-Modal Pre-training aims to effectively leverage paired and unlabeled multi-modal data from local clients to learn a generalizable model $\hat{f}$.

Multi-modal pre-training utilizes paired data from multiple modalities to learn model $\hat{f}$ that can well represent the samples. In this paper, we consider the pre-training task on image and text modalities. We focus on a classic schema in vision-language pre-training, where the model consists of feature encoders w.r.t. X and Y modalities, and both encoders project their inputs into a shared representation space $\mathcal{Z}$. For example, a good pre-trained model that can encode an image of a running dog and its text description "a photo of running dog" to a shared representation space $\mathcal{Z}$, which is called *cross-modal alignment* in (Castrejon et al., 2016; Gao et al., 2024). Suppose the quality of the representation space of pre-trained models can be measured by a loss objective $\mathcal{R} : \mathbb{R}^{d \times d} \to \mathbb{R}$ (e.g., mutual information between representations X and Y with dimension d), federated multi-modal pre-training aims to minimize $\mathbb{E}_{(x,y)) \sim D_{\mathcal{T}}}[\mathcal{R}(\hat{f}(x, y))]$ on the testing dataset $D_{\mathcal{T}}$.

In federated setting, $\hat{f}$ is aggregated from $\hat{f}_i$, which are learned by minimizing $\mathbb{E}_{(x,y)) \sim D_i}[\mathcal{R}(\hat{f}_i(x, y))]$ during local training. $\hat{f}_i$ may capture client-specific information that may not be generalizable across client dataset domains and will affect the performance of the aggregated $\hat{f}$ if local datasets are heterogeneous.

Table 1 provides a comparison of the most similar previous works, highlighting the distinctions between their tasks and ours, as well as the technical differences between their approaches and ours.

4 Method

4.1 Global Constrained Local Training Objective

During local training, the pre-trained model would capture client-specific information that can not generalize to other data domains, as shown in Sec. 5.2. Here, we will provide an in-depth analysis of this phenomenon and propose a global constraint term to alleviate it.

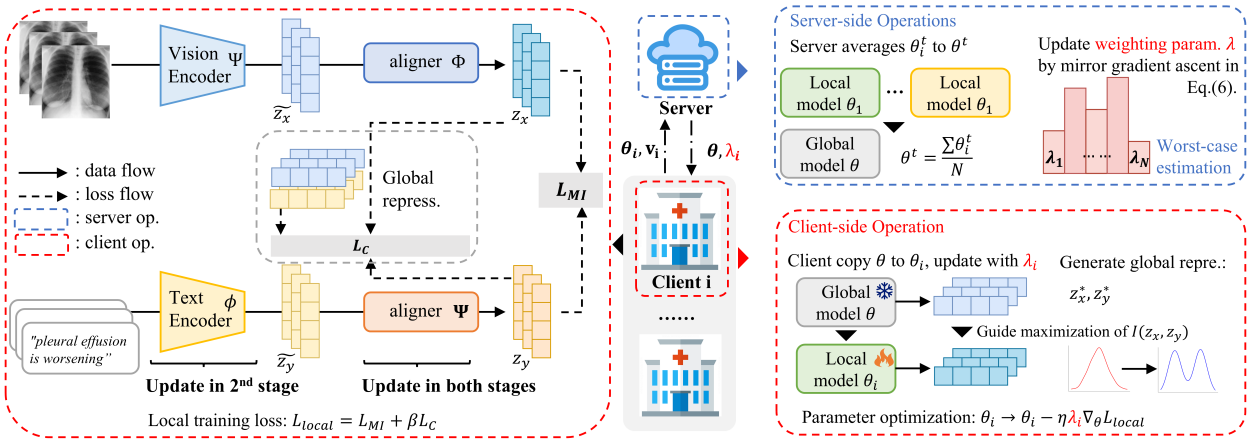

Figure 2: Our proposed FedDRA method follows a two-stage pre-training schema. Alignment modules are first trained in the first stage. Then we jointly update the whole model in the second stage. During local training, we use frozen copies of the server-aggregated models to obtain global representations z_x^* and z_y^*, to regularize local training. After each round of local training, the server estimates the current worst-case distribution and updates the λ_i, which are then sent to each client to adjust the local update step-size.

In classical vision-language pre-training setting, the vision-language model f is composed of two encoders, $f_\psi : \mathcal{X} \rightarrow \mathcal{Z}$ for image modality X, and model $f_\phi : \mathcal{Y} \rightarrow \mathcal{Z}$ for text modality Y. Given an image-text paired data (x, y), these models project the input to representations $z_x = f_\psi(x), z_y = f_\phi(y)$. The goal of vision-language pre-training is to learn a cross-modal alignment from unpaired data, and thus obtain a generalizable representation space, where image representation z_x and text representations z_y are well-aligned. It could be viewed as maximizing the mutual information of representations of the two modalities (Su et al., 2023). Therefore, we can measure the cross-modal alignment degree with a loss mutual-information-based loss objective $\mathcal{R}$, which is approximated by InfoNCE (Liu et al., 2021; Lu et al., 2024) in this paper.

Current multi-modal pre-training methods often encourage the model to maximize mutual information $I(z_x, z_y)$ of the training pairs, neglecting the potential data heterogeneity problem in the federated learning scenarios. As the distribution D_i varies across clients, each client dataset corresponds to a distinct optimal model f_i, which is induced from the distribution of the client dataset. For local training in client i, given that only $(x, y) \sim \mathcal{D}_i$ are available, the locally learned encoders $\hat{f}_i$ tends to move towards f_{ψ_i} and f_{ϕ_i}, and might capture some harmful client-specific information. In federated pre-training, the model is expected to capture patterns that are transferable across clients and potential testing domains, and client-specific information might harm the model's generalization ability. Therefore, it is crucial to explicitly force the model to learn generalizable knowledge. Previous methods such as FedAvg (McMahan et al., 2017; Lu et al., 2023), which do not account for this distinction, may result in learning biased local models $\hat{f}_i$, and diminish the generalization ability of the aggregated model $\hat{f}$.

Given the distribution $D_\mathcal{T}$ of testing data domain, we aim to minimize the generalization error of the federally learned model on $D_\mathcal{T}$. Formally, let $\mathcal{F}_\psi \subseteq \{f_\psi : \mathcal{X} \rightarrow \mathcal{Z}\}$ to be a hypothesis space defined on input space of image modality X and $\mathcal{F}_\phi \subseteq \{f_\phi : \mathcal{Y} \rightarrow \mathcal{Z}\}$ of text modality $\mathcal{Y}$. Suppose encoders $\hat{f}_\psi \in \mathcal{F}_\psi$ and $\hat{f}_\phi \in \mathcal{F}_\phi$ are the federally learned models, and there exist optimal $f_{\psi_i} \in \mathcal{F}_\psi$ and $f_{\phi_i} \in \mathcal{F}_\phi$ for each data domain D_i. Denote $\epsilon_{g,h}(x) = \|g(x) - h(x)\|_2^2$ to be the error of two models $g(\cdot), h(\cdot)$ on data sample x, $\mathcal{L}$ as the InfoNCE loss. We have an upper bound of the generalization error $\mathcal{R}_T(\hat{f}) = \mathbb{E}_{(x,y)\sim D_\mathcal{T}}[\mathcal{L}(\hat{f}_\psi(x), \hat{f}_\psi(y))]$.

Proposition 1. *Let $\{D_i, f_{\psi_i}, f_{\phi_i}\}_{i=1}^N$ and $D_\mathcal{T}, f_{\psi_\mathcal{T}}, f_{\phi_\mathcal{T}}$ be the distributions and optimal encoders for each client data domains and the testing domain, respectively. Given mixed weights $\{w_i\}_{i=1}^N$, $\sum_{i=1}^N w_i = 1$, $w_i \geq 0$, federally learned model $\hat{f}$, and temperature τ in InfoNCE loss. The generalization error $\mathcal{R}_T(\hat{f})$ follows:*

$$\mathcal{R}_T(\hat{f}) \leq \mathbb{E}_{(x,y)\sim\mathcal{D}_\mathcal{T}} \alpha_i \cdot w_i \cdot \left[\epsilon_{\hat{f_\psi}, \hat{f_{\psi_i}}}(x) + \epsilon_{f_{\psi_i}, \hat{f_{\psi_i}}}(x) + \epsilon_{\hat{f_\phi}, \hat{f_{\phi_i}}}(y) + \epsilon_{f_{\phi_i}, \hat{f_{\phi_i}}}(y) + C_i\right].$$

where C_i, α_i are client-specific constants.

Here, $\epsilon_{f_{\psi_i}, f_{\hat{\psi}_i}}(x)$ and $\epsilon_{f_{\phi_i}, f_{\hat{\phi}_i}}(x)$ measure the discrepancy between the locally trained models and the optimal models of the local data domain. These discrepancies are minimized during local training, leading the local models to learn client-specific information. Another two terms $\epsilon_{f_{\hat{\psi}}, f_{\hat{\psi}_i}}(x)$ and $\epsilon_{f_{\hat{\phi}}, f_{\hat{\phi}_i}}(x)$ capture the discrepancy between the server-aggregated models and the locally trained models. Minimizing these terms can not only help reduce the upper bound of the generalization error, but also encourage the local models to learn patterns that generalize well to unseen testing data domains.

Motivated by this, we can directly optimize the terms $\epsilon_{f_{\hat{\psi}}, f_{\hat{\psi}_i}}(x)$ and $\epsilon_{f_{\hat{\phi}}, f_{\hat{\phi}_i}}(x)$ to encourage the models to learn generalizable features. By projecting the inputs x and y through $\hat{f}_{\hat{\psi}}$ and $\hat{f}_{\hat{\phi}}$, we obtain the global representations z_x^* and z_y^*, respectively. Therefore, the constraint loss can be defined as $L_C = ||z_x - z_x^*||^2 + ||z_y - z_y^*||$. The total loss objective for local training could be:

$$L_{local} = L_{MI} + \mu L_C, \tag{1}$$

where μ is the hyper-parameter that adjusts the constrained degree, L_{MI} is the pre-training loss term based on infoNCE loss (e.g., image-text contrastive loss).

4.2 Two-Stage Alignment For Mitigating the Deep Layer Bias

Furthermore, as discussed in Sec. 5.2, deep layers that contain biased client-specific information can impede pre-trained model to learn generalizable representations. This observation is similar to the findings in the supervised federated learning domain (Legate et al., 2024), where training from a better initialized last layer, can less capture biased information in client local datasets. Motivated by this, we model the deep layers of the encoding functions of modality X, Y as alignment modules f_Ψ, f_Φ, and aim to obtain generalizable alignment modules. In practice, we add additional blocks as the alignment module for simplicity, instead of dividing each encoder to two separate parts. For a data flow of an input pair (x, y), the image encoder and text encoder f_ψ, f_ϕ firstly take x, y to obtain intermidiate features $\tilde{z}_x, \tilde{z}_y$, respectively. Then the corresponding alignment modules f_Ψ, f_Φ will project $\tilde{z}_x, \tilde{z}_y$ to aligned final representation z_x, z_y.

To mitigate the negative impact of the biased alignment module on the generalization ability of the pre-trained model, we first train generalizable alignment modules f_Ψ and f_Φ, and then use them to enhance the training of the feature encoders f_ψ and f_ϕ. To encourage alignment modules learn to extract general features, in the first stage, we train them with frozen feature encoders using the learning objective Eq.(1). Since these feature encoders are less biased from client-specific information, alignment modules are encouraged to learn unbiased mappings from $\tilde{z}_y$ to z_y and $\tilde{z}_x$ to z_x. Then in the second stage, we train both the alignment module and feature encoders to enhance their capability of extracting medical features, with the same learning objective as the first stage. The complete pipeline is illustrated in Fig. 2.

4.3 Learning Robust Cross-Modal Alignment via Distributionally Robust Optimization

In real-world scenarios, $\mathcal{D}_\mathcal{T}$ is unknown during pre-training and is typically out-of-distribution. A common approach to address this issue is to assume the distribution of the testing data domain is near the distribution of overall training data (Rahimian & Mehrotra, 2019; Levy et al., 2020), and construct a set $\mathcal{Q}^\mathcal{C}$ that covers potential testing distributions. By optimizing the model's parameter θ over this set with the loss objective $\mathcal{R}(\theta) := \sup_{Q \in \mathcal{Q}^\mathcal{C}} \left\{ \mathbb{E}_{(x,y)\sim Q} \left[L_{\text{local}}(\theta, x, y) \right] \right\}$, we can pre-train a model that generalizes well to the whole set of potential testing distributions. Here, the maximum loss implies a worst-case distribution in $\mathcal{Q}^\mathcal{C}$, where the pre-trained model performs the worst in aligning the two modalities.

Inspired by this, we aim to optimize the loss objective on the worst-case distribution, and introduce the Distributionally-Robust-Optimization (DRO) to our federated multi-modal pre-training task. DRO first construct a family of testing distributions $\mathcal{Q}^\mathcal{C}$ as shown in Sec. 3, and optimize the model's performance on the worst-case distribution, where the model performs the poorest among distributions in $\mathcal{Q}^\mathcal{C}$. However, during federated learning, the server has no access to the distribution of the entire data. Motivated by (Zhang et al., 2023), we introduce a de-centralized form of the DRO problem. The optimization object could be

Algorithm 1 FedDRA for Federated V-L Pre-training

Input: Init. encoder param. $\theta^{(0)}$; Init. alignment module param. $\Theta^{(0)}$; Client weight $\{w_i^{(0)}\}_{i=1}^N$; Uncertainty set radius ρ; Client datasets $\{D_i\}_{i=1}^N$; Local iteration steps E.
for $r = 0, \ldots, R-1$ **do**
  Server broadcasts $\Theta^{(rE)}, \theta^{(rE)}, \theta^{(rE)*}, \{w_i^{(r)}\}_{i=1}^N$
  for client $i = 1, \ldots, N$ **do**
    Utilize Algorithm 2 on D_i to get $\Theta_i^{(r+1)E}$
  end for
  Server computes: $\Theta^{(r+1)E} = \frac{1}{N}\sum_{i=1}^N \Theta_i^{(r+1)E}$
  Server broadcasts $\Theta^{(r+1)E}$
  for client $i = 1, \ldots, N$ **do**
    Utilize Algorithm 3 on D_i to get $\theta_i^{(r+1)E}$
  end for
  Server computes: $\theta^{((r+1)E)} = \frac{1}{N}\sum_{i=1}^N \theta_i^{((r+1)E)}$
  for client $i = 1, \ldots, N$ **do**
    Compute loss $v_i^{(r+1)}$
    of model $(\Theta_i^{(r+1)E}; \theta_i^{(r+1)E})$ on D_i
  end for
  for client $i = 1, \ldots, N$ **do**
    $w_i^{(r+1)} = Proj_{\mathcal{Q}i}\left(w_i^{(r)} e^{\gamma v_i^{(r+1)}} / \sum_{i=1}^N w_i^{(r)} e^{\gamma v_i^{(r+1)}}, \rho\right)$
  end for
end for
return $(\Theta^{(RE)}; \theta^{(RE)})$

Algorithm 2 1st Stage Training

Input: Learning rate η;Update stepsize γ; Local iteration steps E; Param. θ^{rE}, Θ^{rE}; Weight w_i
Set $(\theta_i^{rE}; \Theta_i^{rE}) = (\theta^{rE}; \Theta^{rE})$
Set $(\theta_i^{rE*}; \Theta_i^{rE*}) = (\theta^{rE}; \Theta^{rE})$
for $t = rE, \ldots, (r+1)E-1$ **do**
  Sample ξ_i^t from D_i uniformly
  $\theta_i^{t+1*} = \theta_i^{t*} -$
  $w_c^{r+1}\eta\nabla_\Theta L_{dro}(\Theta_i^t; \theta_i^t; \Theta^{(r+1)E*}; \theta^{(r+1)E*}; \xi_i^t)$
end for

Algorithm 3 2nd Stage Training

Input: Learning rate η; Update stepsize γ; Local iteration steps E; Param. $\theta^{rE}, \Theta^{(r+1)E}$; Weight w_i
Set $(\theta_i^{rE}; \Theta_i^{(r+1)E}) = (\theta^{rE}; \Theta^{(r+1)E})$
Set $(\theta_i^{rE*}; \Theta_i^{(r+1)E*}) = (\theta^{rE}; \Theta^{(r+1)E})$
for $t = rE, \ldots, (r+1)E-1$ **do**
  Sample $\xi_i^{(t)}$ from D_i uniformly
  $(\theta_i^{t+1}; \Theta_i^{t+1}) = (\theta_i^t; \Theta_i^t) - w_c^{r+1}$
  $\eta\nabla_{(\theta_i;\Theta_i)} L_{local}(\theta_i^t; \Theta_i^t; \Theta^{(r+1)E*}; \theta^{(r+1)E*}; \xi_i^t)$
end for
The client sends $\theta_i^{(r+1)E}, v_c^r$ to the server

written as:

$$\sup_{\lambda \in \Delta_{|\mathcal{C}|-1}} \left\{ R(\theta, \lambda) := \sum_i \lambda_i R_i(\theta) \text{ s.t. } D_f(|\mathcal{C}| \cdot \lambda || (1, 1, \ldots, 1)) \leq \rho \right\}, \tag{2}$$

where $R_i(\theta) := \mathbb{E}_{(x,y)\sim D_i}[L_{local}(\theta; (x, y))]$ is the empirical risk on client data D_i, ρ is the uncertain radius as mentioned in Sec. 3.

Then, we can optimize Eq. 2 by alternatively optimize the weights λ and model parameters θ. Specifically, we optimize the parameter θ_i of local models on each iteration t by $\theta_i^{(t+1)} = \theta_i^{(t)} - \eta\lambda_i^t \nabla_\theta L_{local}$, where η is the learning rate. Following the mirror gradient ascent of weight proposed in (Zhang et al., 2023), we update λ with $\lambda^{t+1} = \frac{\lambda_i^t e^{\gamma v_i^t}}{\sum_{i=1}^{|\mathcal{C}|} \lambda_j^t e^{\gamma v_j^t}}$. Then we compute λ^{t+1} by projecting $\tilde{\lambda}^{t+1}$ into the set $\{\lambda : D_f(|\mathcal{C}|) \cdot \|\lambda\| \|(1, 1, \ldots, 1)\| \leq \rho\}$ to fit the constraints of the uncertainty set. In practice, we update λ after the local training of each communication turn.

We apply the proposed DRO in both the first stage of training the alignment module and the second stage of training the feature encoder. In both stages, we use the same objective, L_{local}, to encourage the model to learn generalizable information and mitigate the impact of data heterogeneity on maximizing the mutual information $I(z_x, z_y)$. The key difference between the two stages is, in the first stage, the optimization target θ in Eq. 2 corresponds to the parameters of the alignment modules Φ and Ψ, whereas in the second stage, θ represents the parameters of the feature encoders ϕ, ψ and alignment modules Φ, Ψ. The pseudo-code of the whole algorithm can be seen in Algorithm 1.

5 Experiment

5.1 Experiment Setting

We focus on adapting medical vision-language pre-training methods to heterogeneous federated learning settings. We employ the framework of image-text contrastive learning with two modality-specific encoders, a fundamental design in multi-modal pre-training. We use vision-language pre-training tasks on Chest X-ray datasets and ophthalmology image datasets to evaluate the effectiveness of our FedDRA method.

5.1.1 Experiment Set-up of Pre-training on Chest X-Ray datasets

Pre-training setup. Following (Wang et al., 2022), we utilize the MIMIC-CXR (Bigolin Lanfredi et al., 2022) dataset for medical vision-language pre-training. Following (Yan et al., 2023), we employ the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to divide the MIMIC-CXR dataset based on disease labels to construct 5 heterogeneous client datasets. We set the heterogeneity degree in the LDA algorithm to be 1. Each divided dataset consists of train splits and test splits based on the notation of the MIMIC-CXR. We use the train split for pre-training, and test split to evaluate pre-trained model's image-text retrieval performance. Here, we only divide the raw data into 5 subsets, because vision-language pre-training requires a large batchsize and is data-consuming, thus we need to guarantee each client has $10k$ to $50k$ paired data.

For main experiments, we set the number of communication turns to 25, and randomly sample 50 batches of data for local training at each turn. Here, we choose a relatively small number of communication turns compared to classical supervised federated learning. That's because VLP needs large local optimization steps per turn to extract cross-modal alignment.

Downstream tasks. Following (Wang et al., 2022), we conduct the following downstream tasks to evaluate the transferability and generalization ability of the pre-trained model. **(1) Few-shot classification.** We test their performance on multiple image classification benchmarks RSNA Pneumonia Detection (RSNA) (Shih et al., 2019), and Covidx (Wang et al., 2020). We fine-tune our pre-trained model with an additional linear layer on 1%, 10% percentage of the training dataset, and evaluate the classification accuracy. **(2) Medical image segmentation.** We conduct medical image segmentation experiments on the RSNA (Wang et al., 2020) benchmark. We freeze the encoder and fine-tune a U-Net decoder using 1%, 10% of the training data, and then use the Dice score for evaluation. The datasets we have used for the fine-tuning based tasks are out-of-distribution, so that we can evaluate the transferability of the pre-trained model. **(3) Image retrieval.** We utilize the test splits of client datasets for evaluation, these datasets are unseen in pre-training, and can be viewed as in-domain samples. We report the top-1 recall accuracy and top-5 recall accuracy.

5.1.2 Experiment Set-Up of Pre-training on Ophthalmology Datasets

Pre-training setup. We conduct vision-language multi-modal pre-training using retinal image datasets from different institutes to simulate a more real-world setting. These retinal datasets are from different institutions of low-income and high-income countries, and are highly heterogeneous real-world scenes. Specifically, we utilize MESSIDOR (Decencière et al., 2014) from France and BRSET (Nakayama et al., 2023) from Brazil as pre-training datasets, and assign them to two clients. These datasets include both images and tabular EHR records indicating Diabetic Retinopathy (DR) status and edema risk. For implementation, we transform these tabular data into text captions.

Downstream tasks We evaluate the transferability of the models on few-shot classification tasks using the MBRSET (Nakayama et al.) dataset. Unlike the pre-training datasets, MBRSET was collected by portable devices, resulting in a significant distribution shift. We perform few-shot classification tasks on diabetic retinopathy and edema status using this dataset. We fine-tune the model with an additional linear layer on 10%, 20% and 100% of the training data, and report classification accuracies.

5.1.3 Backbones and Baselines

We focus on enabling medical multi-modal pre-training methods to be applied to heterogeneous federated learning scenes. We have considered the generalization ability of our method on different backbone VLP methods, and adopted contrastive-learning-based methods: simple language-image contrastive alignment (ConVIRT) (Zhang et al., 2022; Radford et al., 2021), global-local language-image contrastive alignment(GLoRIA) (Huang et al., 2021), and Multi-Granularity Cross-modal Alignment (MGCA) (Wang et al., 2022). All of the loss objectives of these pre-training methods contain a contrastive loss term, which act as the infoNCE loss to maximize the mutual information between two modalities. And we take this loss term for computing the client weights in the DRO part.

For baseline federated learning strategies, we have adapted FedMAE (Yan et al., 2023), FedEMA (Zhuang et al., 2022), FedMOON (Li et al., 2021b), FedX (Han et al., 2022), FedU (Zhuang et al., 2021), FedL-

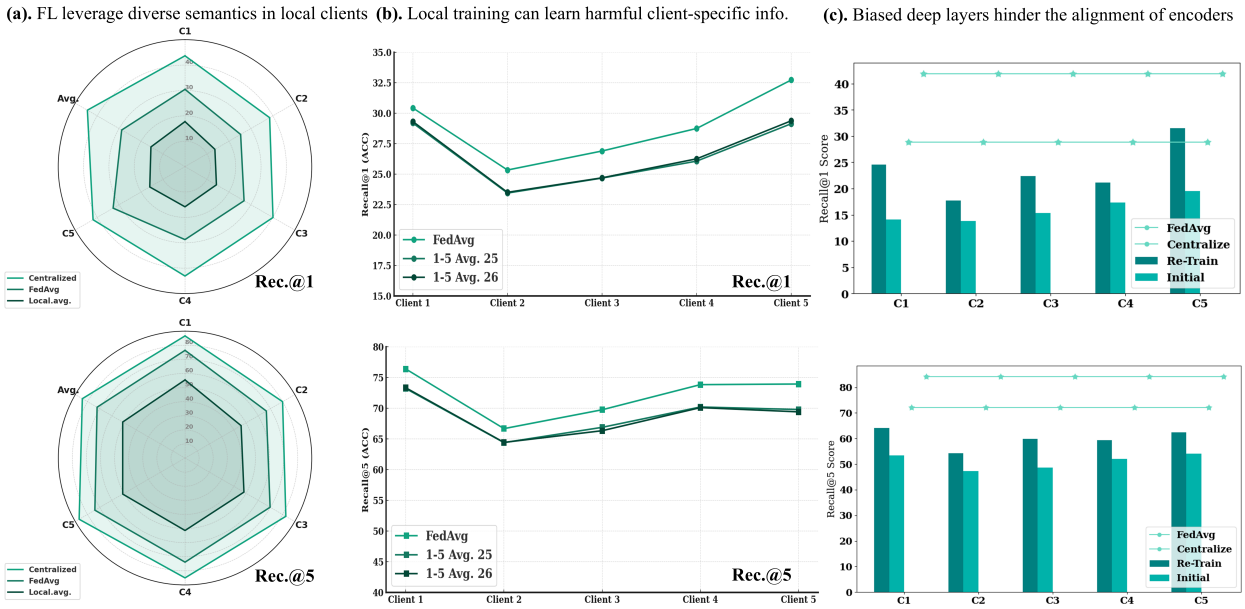

Figure 3: (a). The comparison of retrieval acc. on each client denoted as $\{C_i\}_{i=1}^{5}$, of centralized, FedAvg, and averaged acc. of decentralized pre-trained models. (b). The performance of the server model after 25 com. turns and the averaged performance of corresponding client models after 25 and 26 com. turns, on each client denoted as $\{\text{Client}_i\}_{i=1}^{5}$. (c). The averaged acc. on each client. We show the acc. of centralized and FedAvg pre-trained baselines, and de-centralized pre-trained models shown as $\{C_i\}_{i=1}^{5}$ retrained on the union of training splits of client datasets.

DAWA (Rehman et al., 2023) for comparison. These are self-supervised learning methods in the federated learning domain which also focus on tackling the data heterogeneity. For basic federated learning baselines, we consider simple averaging (FedAvg) (McMahan et al., 2017), decentralized training, and centralized training. For baselines pre-trained in Local strategy, we report the averaged performance of the local models.

For fair comparisons, we re-implemented all baseline methods using the same backbones. To adapt uni-modal self-supervised learning baselines to our setting, we added an image-text contrastive loss, applying the same hyperparameters as in our method for consistency. We use ViT-base (Dosovitskiy et al., 2020) as the vision encoder and Bert-base (Devlin et al., 2018) as the text encoder, with input pre-processing following (Wang et al., 2022). Additionally, we employ an extra transformer block from ViT-base and Bert-base as the alignment module for vision and language, respectively.

5.2 Empirical Finding

In this section, we will demonstrate our key empirical findings for federated multi-modal pre-training under heterogeneous client datasets, which actually motivates us to propose our FedDRA. We conduct experiments on the image-text retrieval task, which reflects the ability to maximize the mutual information and learn cross-modal alignment. In the following studies, we mainly compare performances of naive federated (FedAvg) pre-trained model, decentralized pre-trained model, and centralized pre-trained model.

Federated learning enhances pre-training by leveraging more samples in a privacy-preserved manner, while data heterogeneity can affect the effectiveness of the FedAvg. Figure 3(a) presents the retrieval accuracies of the models under consideration. Despite the heterogeneity of local datasets, the FedAvg strategy significantly outperforms the decentralized pre-training approach. However, the centralized pre-trained model remains an upper bound, indicating substantial room for improvement.

Local training can learn harmful client-specific information, degrading the performance of the pre-trained model. After several communication turns, re-training the aggregated server model on local datasets may lead to a performance drop, as shown in Figure 3(b). We considered a set of models retrained

on a server model with local datasets respectively. The server model is learned through 25 communication turns. Compared to the starting server model, the averaged accuracy of local retrained models is significantly lower. This degradation may be because local training focus on learning domain-specific information in the late communication rounds, which would affect the aggregated model's overall performance.

Decentralized pre-trained deep layers can hinder the learning of a generalizable feature extractor. We re-trained the first four shallow layers of the decentralized pre-trained model on the combined local datasets. While this led to some performance improvements, a significant gap still remains compared to the FedAvg pre-trained baselines, as shown in Figure 3(c). This gap indicates that the biased frozen deep layers prevent the model from learning more diverse semantics from the combined dataset. We hypothesize that these deep layers may contain biased, client-specific information, which obstructs the cross-modal alignment process. Our findings align with observations (Legate et al., 2024) in supervised federated learning.

Overall, from empirical findings, we conclude that federated multimodal pre-training is sensitive to data heterogeneity, and simply averaging local model weights does not solve this issue essentially. Furthermore, performance is closely tied to the generalization ability of the final layers in pre-trained models. Thus, we are motivated to propose our method.

Table 2: Downstream task performance. We report the few-shot classification accuracy on Covid and RSNA, the in-domain retrieval accuracy, and the Dice score for segmentation on RSNA.

Strategy	**Backbone**	**RSNA (cls.)**		**Covid (cls.)**		**RSNA (seg.)**		**In-domain Image-Text Retrieval**			
		1%	10%	1%	10%	1%	10%	**Rec.@1**	**Rec.@5**	**Wst.@1**	**Wst.@5**
FedEMA	ConVIRT	82.8	83.1	79.2	86.5	70.9	73.6	24.0	67.0	21.9	62.4
FedMOON	ConVIRT	82.5	83.2	77.8	89.2	69.0	71.3	27.8	70.9	25.3	67.2
FedAvg	ConVIRT	83.1	83.3	78.0	88.5	69.6	71.5	28.8	72.1	25.3	66.7
FedDRA (Ours)	ConVIRT	**83.2**	**83.7**	**81.0**	**90.3**	**71.7**	**74.1**	**30.2**	**73.2**	**27.0**	**68.9**
FedX	GLoRIA	82.7	83.4	78.3	88.5	71.0	72.1	28.5	72.2	25.9	68.0
FedU	GLoRIA	83.0	83.5	78.7	89.3	71.2	72.6	29.2	73.0	27.6	69.5
FedAvg	GLoRIA	83.2	83.3	77.5	89.0	71.4	72.4	29.9	73.8	27.8	69.5
FedDRA (Ours)	GLoRIA	**83.6**	**84.1**	**79.4**	**89.8**	**72.0**	**73.2**	**31.1**	**74.3**	**28.2**	**70.2**
FedLDAWA	MGCA	82.4	83.5	78.1	88.5	70.4	72.6	29.0	73.5	27.0	68.9
FedAvg	MGCA	82.6	83.5	75.8	88.2	70.1	71.4	29.3	73.7	26.8	70.4
FedDRA (Ours)	MGCA	**83.1**	**83.8**	**79.3**	**89.1**	**71.0**	**72.8**	**29.8**	**74.1**	**27.4**	**70.6**

Table 3: We have conducted ablation experiments to verify the effectiveness of our key technical designs. We report the few-shot classification accuracy and the retrieval accuracy.

Two-stage	**Global Constraint**	**DRO-Weighing**	**Covid (cls.)**		**RSNA (cls.)**		**In-domain Image-Text Retrieval**			
			1%	10%	1%	10%	**Rec.@1**	**Rec.@5**	**Wst.@1**	**Wst.@5**
	✓	✓	83.0	83.4	80.5	89.6	29.4	72.7	26.2	68.1
✓		✓	82.8	83.0	79.8	88.6	28.3	71.9	26.0	67.9
✓	✓		82.5	82.9	80.2	89.2	29.7	72.8	25.6	67.3
✓	✓	✓	**83.2**	**83.7**	**81.0**	**90.3**	**30.2**	**73.2**	**27.0**	**68.9**

5.3 Main Results

Our method learns robust and enriched cross-modal alignment and has better transferability. Table 2 has shown results of downstream tasks, here we utilize the ConVIRT as the backbone pre-training method. In the image-text retrievel task, both average and worst-client accuracies of our method are higher than baseline's, which means our model can capture more robust cross-client features. In the few-shot classification and segmentation, our method beats other baseline strategies on each task, which demonstrate the higher generalization ability of the representation space learned by our method.

Table 2 has shown the performance of adapted self-supervised federated learning methods which focus on single-modality. From the results, we observe that baselines have shown better transferability on visual

downstream tasks, compared to the naive FedAvg strategy. However, in the multi-modal retrieval task, our method beats these baselines by a large margin, which indicates that previous single-modality methods cannot be easily adapted directly for multi-modal data. Furthermore, FedAvg is a strong baseline in multi-modal retrieval tasks compared to other adapted methods, as we observed in the experiments. We conjecture that's because FedAvg only focuses on maximizing the in-domain mutual information, and doesn't introduce additional loss terms which would hurt the learning of enriched cross-modal alignment. However, this may lead to lower generalization ability on few-shot downstream tasks as discussed before.

Table 4: Downstream task performance. We report the few-shot classification accuracy of Diabetic Retinopathy and Edema Risk classification tasks on the MBRSET dataset.

Strategy	**Diabetic Retinopathy (cls.)**			**Risk of Edema (cls.)**		
	10%	20%	100%	10%	20%	100%
Decentralized	78.8	79.7	81.1	91.5	92.5	93.8
FedAvg	79.4	80.2	82.3	92.8	93.6	94.2
FedMAE (Yan et al., 2023)	79.2	80.3	82.0	92.4	93.3	94.0
FedX (Han et al., 2022)	79.5	80.1	81.6	93.0	93.5	94.3
FedU (Zhuang et al., 2021)	79.7	80.5	81.7	92.8	93.4	94.1
FedDRA (Ours)	**80.6**	**81.5**	**83.1**	**93.4**	**94.1**	**94.9**
FedGlobal	81.9	82.6	84.0	94.2	94.7	95.8

Our method can be transferred to multiple multi-modal pre-training methods. Table 2 shows the downstream task performance of the MGCA and GLoRIA backbone pre-training methods when combined with our strategy. Our method has successfully adapted MGCA and GLoRIA to the heterogeneous federated multi-modal pre-training scenario, as demonstrated by the significant improvement in classification and segmentation tasks.

5.4 Analysis Experiments

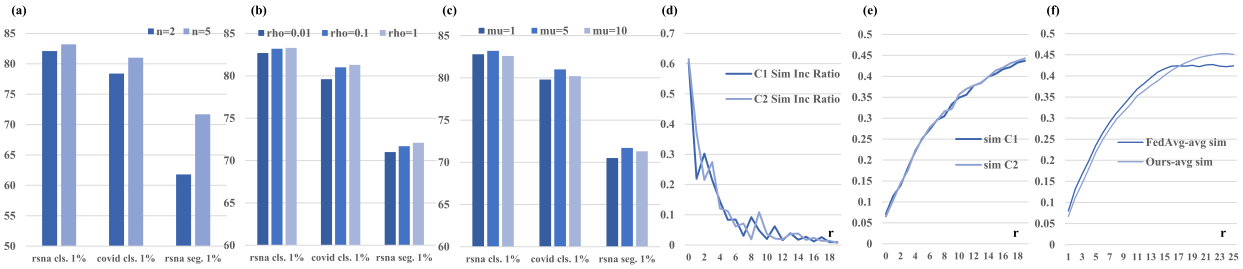

Figure 4: (a) Analysis study on number of clients. (b) Analysis study on uncertainty radius. (c) Analysis study on global constraint degree. (d) Improvement of image-text embedding similarity per.commu turn. (e) Image-text embedding similarity curve of our method. (f) Averaged similarity curves.

The two-stage pre-training strategy and global constraints can enhance the learning of cross-modal alignement. We remove the global constraint loss L_C of our method, and compare the pre-trained model's performances with those of the original version. As shown in Table 3, the downstream performance, particularly the image-text retrieval accuracy, is significantly lower in the modified version. Similarly, we remove the first-stage pre-training on alignment modules to verify the role of the two-stage pre-training strategy. We have found that the first stage pre-training can help learn better cross-modal alignment and achieve high image-text retrieval accuracies, as shown in Table 3.

DRO weighing can reduce the domain gap and improve the downstream performances. We remove the DRO weighing part and compare the performances of the model pre-trained with the original method. As shown in Table 3, removing the DRO-weighing part leads to a large performance drop in few-shot classification performances, thus adding DRO-weighing component can improve the transferability of the pre-trained model. The client-wise worst accuracies of the original model are much higher than those of the modified version. This means DRO has succesfully bridged the gap between local training data and downstream dataset, by optimizing model performance on the constructed uncertain set of distributions.

FedDRA dynamically schedules updating stepsize for each client, and therefore optimizes the worst-case performance. We select two clients C_1, C_2 in the federated pre-training. For each client, we calculate the average cosine similarity between image and text embeddings, using the server-aggregated

Table 5: Results on client datasets with different heterogeneity degrees, across different pre-training methods.

Strategy	α	**Backbone**	**RSNA (cls.)**		**Covid (cls.)**		**RSNA (seg.)**		**Backbone**	**RSNA (cls.)**		**Covid (cls.)**		**RSNA (seg.)**	
			1%	10%	1%	10%	1%	10%		1%	10%	1%	10%	1%	10%
FedAvg	1	ConVIRT	82.2	83.3	78.0	88.5	69.6	71.5	GLoRIA	83.2	83.8	78.5	89.0	71.4	72.4
FedDRA (ours)	1	ConVIRT	83.2	83.7	81.0	90.3	71.7	74.1	GLoRIA	81.8	82.9	78.5	89.6	69.3	72.5
FedAvg	5	ConVIRT	80.7	82.3	77.6	88.2	68.5	71.7	GLoRIA	81.8	82.9	78.5	89.6	69.3	72.5
FedDRA (ours)	5	ConVIRT	81.8	82.9	78.5	89.6	69.3	72.5	GLoRIA	82.2	83.0	78.4	89.4	71.5	72.9
Centralized	-	ConVIRT	83.4	84.6	82.5	92.0	72.6	76.4	GLoRIA	84.0	84.7	82.2	91.8	73.5	73.7

model at each communication turn. In Fig. 4(d) and Fig. 4(e), we plot curves of these similarities, which reflect the cross-modal alignment degree. At each communication turn, when the similarity of a client is relatively high, its similarity in next turn would get a smaller improvement. That's because our FedDRA can assign higher updating stepsize to client where the cross-modal alignment is less extracted by the model.

Our FedDRA can alleviate over-fitting client-specific information, and learn better cross-modal alignment For our FedDRA and the FedAvg method, we plot the cosine-similarities of text and image embedding averaged across clients at each communication turn. As shown in Fig.4(f), FedAvg requires fewer communication rounds to converge, but results in fluctuations after certain communication turns. This aligns with findings in Sec.5.2, where local retraining a model that are trained after multiple communication rounds, would introduce harmful client-specific information and distort the learned representation space. In contrast, our FedDRA gradually extract cross-modal alignment from local training in a distributionally robust manner, and learns a stronger representation space.

Analysis on global constraint hyper-parameter μ. As shown in Fig. 4(c), a larger μ encourages federated pre-training to enhance performance on less optimized client data domains, leading to smaller disparity on image-text retrieval performance on each client domain. As we increase μ from 1 to 5, the downstream performance consistently increases. However, a excessively large μ can decrease overall performance.

A larger uncertainty radius ρ improves transferability in downstream tasks. Fig. 4(b) shows the downstream performance of models pre-trained with different uncertainty radii in the DRO process. As larger ρ would bring higher performance in few-shot classification and segmentation tasks on out-of-domain datasets. We also observed that a smaller ρ better supports cross-modal alignment learning, achieving better image-text retrieval performance on in-domain datasets, as shown in Table 8 in the Appendix. This is because the larger uncertainty radius would incorporate more potential out-of-distribution cases, which can enhance the model's transferability.

Robust check on heterogeneity degree of client datasets. We changed the α which adjusts the heterogeneity degree of the LDA allocated client datasets, to check the robustness of our method under different heterogeneity degree. As shown in Table 5, our method consistently enhance pre-training methods' performances under client datasets with different heterogeneity degrees.

Robust check on number of clients. We adjusted the number of clients involved in federated pre-training. As shown in Fig. 4(a), increasing the number of clients introduces greater diversity, which can enhance the downstream performance of the pre-trained model.

6 Conclusion

Data limitation is a long-standing problem in the multi-modal learning domain. Despite federated learning can leveraging datasets from multiple sources while guaranteeing privacy issues, its performance would be damaged by data heterogeneity. Inspired by our empirical findings on the impact of heterogeneity on federated multi-modal learning, we propose the FedDRA framework to mitigate heterogeneity for federated medical vision-language pre-training. The effectiveness of our method has been verified by comprehensive experiments. While introducing representation from other clients might bring larger improvement, we still consider the most privacy-preserved setting where representations are not transmissible. Further work could explore how to introduce diversity of multi-modal pre-training data while keeping local data private.

References

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15016–15027, 2023.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.

Ricardo Bigolin Lanfredi, Mingyuan Zhang, William F Auffermann, Jessica Chan, Phuong-Anh T Duong, Vivek Srikumar, Trafton Drew, Joyce D Schroeder, and Tolga Tasdizen. Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. *Scientific data*, 9(1):350, 2022.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Giacomo Capitani, Federico Bolelli, Angelo Porrello, Simone Calderara, and Elisa Ficarra. Clusterfix: A cluster-based debiasing approach without protected-group supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4870–4879, 2024.

Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2940–2949, 2016.

Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pp. 1542–1553. PMLR, 2020.

Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 679–689. Springer, 2022.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pp. 2089–2099. PMLR, 2021.

Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, John-Richard Ordóñez-Varela, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, pp. 231–234, 2014.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. *Advances in neural information processing systems*, 33:15111–15122, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Enwei Zhang, Ke Li, Jie Yang, Wei Liu, and Xing Sun. Softclip: Softer cross-modal alignment makes clip stronger. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1860–1868, 2024.

Avishek Ghosh, Justin Hong, Dong Yin, and Kannan Ramchandran. Robust federated learning in a heterogeneous environment. *arXiv preprint arXiv:1906.06629*, 2019.

Peixuan Han, Zhenghao Liu, Zhiyuan Liu, and Chenyan Xiong. Distributionally robust unsupervised dense retrieval training on web graphs. *arXiv preprint arXiv:2310.16605*, 2023.

Sungwon Han, Sungwon Park, Fangzhao Wu, Sundong Kim, Chuhan Wu, Xing Xie, and Meeyoung Cha. Fedx: Unsupervised federated learning with cross knowledge distillation. In *European Conference on Computer Vision*, pp. 691–707. Springer, 2022.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3942–3951, 2021.

Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10143–10153, 2022.

Colton Ladbury, Arya Amini, Ameish Govindarajan, Isa Mambetsariev, Dan J Raz, Erminia Massarelli, Terence Williams, Andrei Rodin, and Ravi Salgia. Integration of artificial intelligence in lung cancer: Rise of the machine. *Cell Reports Medicine*, 2023.

Gwen Legate, Nicolas Bernier, Lucas Page-Caccia, Edouard Oyallon, and Eugene Belilovsky. Guiding the last layer in federated learning with pre-trained models. *Advances in Neural Information Processing Systems*, 36, 2024.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019.

Jingtao Li, Lingjuan Lyu, Daisuke Iso, Chaitali Chakrabarti, and Michael Spranger. Mocosfl: enabling cross-client collaborative self-supervised learning. In *The Eleventh International Conference on Learning Representations*, 2022a.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022b.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10713–10722, 2021b.

Chen Liu, Yanwei Fu, Chengming Xu, Siqian Yang, Jilin Li, Chengjie Wang, and Li Zhang. Learning a few-shot embedding model with contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 8635–8643, 2021.

Jiashuo Liu, Zheyan Shen, Peng Cui, Linjun Zhou, Kun Kuang, and Bo Li. Distributionally robust learning with stable adversarial training. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Siyu Lu, Zheng Liu, Tianlin Liu, and Wangchunshu Zhou. Scaling-up medical vision-and-language representation learning with federated learning. *Engineering Applications of Artificial Intelligence*, 126:107037, 2023.

Yiwei Lu, Guojun Zhang, Sun Sun, Hongyu Guo, and Yaoliang Yu. f-micl: Understanding and generalizing infonce-based contrastive learning. *arXiv preprint arXiv:2402.10150*, 2024.

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*, pp. 1273–1282. PMLR, 2017.

Paolo Mendogni, Jacopo Vannucci, Marco Ghisalberti, Marco Anile, Beatrice Aramini, Maria Teresa Congedo, Mario Nosotti, Luca Bertolaccini, on behalf of the Italian Society for Thoracic Surgery (endorsed by the Italian Ministry of Health) Collaborators of the Pneumothorax Working Group Collaborators of the Pneumothorax Working Group, Ambra Enrica D'Ambrosio, et al. Epidemiology and management of primary spontaneous pneumothorax: a systematic review. *Interactive cardiovascular and thoracic surgery*, 30(3):337–345, 2020.

Luis Filipe Nakayama, Mariana Goncalves, L Zago Ribeiro, Helen Santos, Daniel Ferraz, Fernando Malerbi, Leo Anthony Celi, and Caio Regatieri. A brazilian multilabel ophthalmological dataset (brset). *PhysioNet https://doi. org/10*, 13026, 2023.

Luis Filipe Nakayama et al. mbrset, a mobile brazilian retinal dataset.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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 *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

Yasar Abbas Ur Rehman, Yan Gao, Pedro Porto Buarque De Gusmão, Mina Alibeigi, Jiajun Shen, and Nicholas D Lane. L-dawa: Layer-wise divergence aware weight aggregation in federated self-supervised visual representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16464–16473, 2023.

Khaled Saab, Sarah Hooper, Mayee Chen, Michael Zhang, Daniel Rubin, and Christopher Ré. Reducing reliance on spurious features in medical image classification with spatial specificity. In *Machine Learning for Healthcare Conference*, pp. 760–784. PMLR, 2022.

George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

Weijie Su, Xizhou Zhu, Chenxin Tao, Lewei Lu, Bin Li, Gao Huang, Yu Qiao, Xiaogang Wang, Jie Zhou, and Jifeng Dai. Towards all-in-one pre-training via maximizing multi-modal mutual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15888–15899, 2023.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35:33536–33549, 2022.

Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific reports*, 10(1):19549, 2020.

Rui Yan, Liangqiong Qu, Qingyue Wei, Shih-Cheng Huang, Liyue Shen, Daniel Rubin, Lei Xing, and Yuyin Zhou. Label-efficient self-supervised federated learning for tackling data heterogeneity in medical imaging. *IEEE Transactions on Medical Imaging*, 2023.

Fengda Zhang, Kun Kuang, Long Chen, Zhaoyang You, Tao Shen, Jun Xiao, Yin Zhang, Chao Wu, Fei Wu, Yueting Zhuang, et al. Federated unsupervised representation learning. *Frontiers of Information Technology & Electronic Engineering*, 24(8):1181–1193, 2023.

Jianqing Zhang, Yang Hua, Jian Cao, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Eliminating domain bias for federated learning in representation space. *Advances in Neural Information Processing Systems*, 36, 2024a.

Yilan Zhang, Yingxue Xu, Jianqi Chen, Fengying Xie, and Hao Chen. Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. *arXiv preprint arXiv:2401.01646*, 2024b.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pp. 2–25. PMLR, 2022.

Weiming Zhuang, Xin Gan, Yonggang Wen, Shuai Zhang, and Shuai Yi. Collaborative unsupervised visual representation learning from decentralized data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4912–4921, 2021.

Weiming Zhuang, Yonggang Wen, and Shuai Zhang. Divergence-aware federated self-supervised learning. *arXiv preprint arXiv:2204.04385*, 2022.

A Implementation Detailed

A.1 Details of MIMIC-CXR

A.1.1 Pre-training setup

Following (Wang et al., 2022) we utilize the MIMIC-CXR (Bigolin Lanfredi et al., 2022) dataset for multi-modal pre-training. This dataset is widely used in the medical multi-modal learning domain, with $227,835$ image-text pairs from $65,379$ patients. Some related works also have imported additional features to image-text pairs to augment the data. However, we only use the image-text pairs for pre-training to make the results and conclusions more generalizable. The MIMIC-CXR dataset is open access, it can be obtained through MIMIC-CXR Access.

During the pre-training, local clients only have access to their highly heterogeneous datasets. To construct the heterogeneous client datasets, following (Yan et al., 2023) we employ the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to divide the MIMIC-CXR dataset into 5 partitions based on a selected sensitive attribute. For implementation, we import the corresponding attribute information of given image-text pairs from the MIMIC-CXR and divide local datasets based on disease category. The disease category is a multi-label binary attribute and is transformed into a multi-class label. That's because the words in the clinical report are highly related to the disease category as illustrated in Fig 5. We set the heterogeneity degree in the LDA algorithm to be 1 for main experiments. For analysis experiments, we also have run experiments on client datasets allocated by LDA with a heterogeneity degree of 5.

Specifically, we select 5 commonly considered diseases Bannur et al. (2023): 'Edema','Pleural Effusion', 'Consolidation', 'Pneumothorax', and 'Pneumonia'. We set the non-NaN value to 1 and then set NaN value to 0 to construct a 5-way binary multi-label. Then we get 2^5-category multi-class label and run LDA on them.

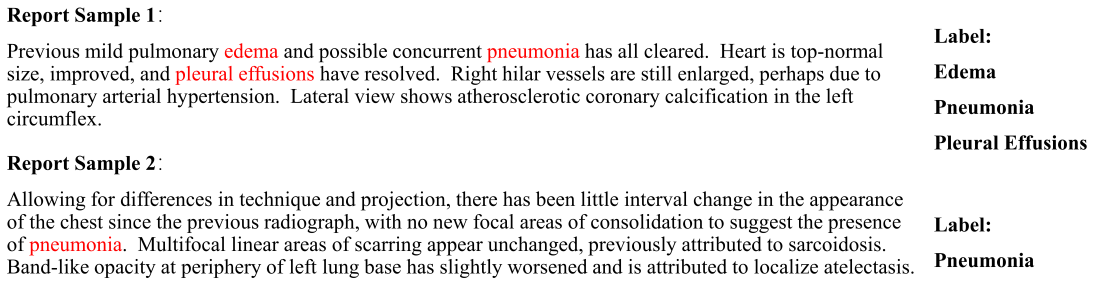

Figure 5: Illustration of the strong connection between latent variable and the text modality.

We divide the MIMIC-CXR into 5 heterogeneous subgroups to construct 5 client datasets. Each divided dataset consists of train splits and test splits based on the notation of the MIMIC-CXR. Our pre-trainings are mainly conducted on $4 \times A40$ or $2 \times A100$. The batch size we have utilized ranged from 288 to 388. We set the learning rate to 2×10^{-5} in main experiments, the number of communications to 25. For our method, we set the uncertainty radius $\rho = 0.1, \mu = 5$ in main experiments. For each communication, we randomly sample 50 batches of data from the client datasets.

A.1.2 Downstream tasks

We evaluate the generalization ability of the pre-trained model through three downstream tasks: few-shot classification, medical image segmentation, and image retrieval.

Few-shot classification. To assess the model's effectiveness on general medical image tasks, we evaluate it on multiple image classification benchmarks: (1) RSNA Pneumonia Detection (RSNA)Shih et al. (2019), where the task is to predict whether an image shows pneumonia. (2) CovidxWang et al. (2020), which includes three categories: COVID-19, non-COVID pneumonia, and normal. We fine-tune our pre-trained

model with an additional linear layer on 1% and 10% of the training dataset and report classification accuracy on these benchmarks.

Medical image segmentation. To explore the model's transferability to fine-grained tasks, we conduct experiments on medical image segmentation using the RSNA Wang et al. (2020) benchmark. Following Wang et al. (2022), we convert RSNA object detection ground truths into segmentation masks. Similar to Huang et al. (2021), we employ a U-Net framework with our pre-trained image encoder as the frozen encoder, while fine-tuning the decoder on 1% and 10% of the training data. The Dice score is used for performance evaluation.

Image retrieval. To verify whether the pre-trained models have captured the semantic alignment between image and text in the pre-training data, we perform an image retrieval task. We test image retrieval performance on the validation splits of the local clients. For each text in a batch of image-text pairs, we calculate similarities with images in the batch, then rank these similarities and retrieve the top-1 and top-5 images. If the corresponding image of the text is in the selected set, it is correctly retrieved. We use top-1 and top-5 recall accuracy to evaluate performance.

A.2 Ophthalmology datasets

A.2.1 Pre-training setup.

We conduct vision-language multi-modal pre-training using retinal image datasets from different institutes. These retinal datasets are from different institutions of low-income and high-income countries, and are highly heterogeneous real-world scenes. Specifically, we utilize MESSIDOR (Decencière et al., 2014) from France and BRSET (Nakayama et al., 2023) from Brazil as pre-training datasets, and assign them to two clients. These datasets include tabular EHR records indicating Diabetic Retinopathy (DR) status and edema risk. We transform tabular data into text captions in the format: "retinal image with {DR status} and {edema risk}" to obtain text prompts. Similar to MIMIC dataset, our pre-trainings on ophthalmology datasets are mainly conducted on $4 \times A40$ or $2 \times A100$. We set the batch size to 100, the number of communications to 20, and the learning rate to 1×10^{-5} in the experiments. For our method, we set the uncertainty radius $\rho = 0.5, \mu = 1$ in main experiments. For each communication, we randomly sample 20 batches of data from the client datasets.

A.2.2 Downstream tasks.

We evaluate the transferability of the models on few-shot classification tasks using the MBRSET (Nakayama et al.) dataset. Unlike the pre-training datasets, MBRSET was collected in low-income areas using portable devices, resulting in a significant distribution shift. We perform few-shot classification tasks on diabetic retinopathy and edema status prediction tasks using this dataset. These are binary classification problems. We fine-tune the model with an additional linear layer on 10%, 20% and 100% of the training data, and report classification accuracies.

B Additional Experiment Results

Federated pre-trained models still show a significant performance gap compared to centralized pre-trained models in multi-modal retrieval tasks. Table 6 shows the performance of models pre-trained in decentralized, FedAvg, centralized federated learning strategies, using different backbone pre-training methods. FedAvg has more effectively extract cross-modal alignment from federally utilizing local datasets, and achieved much better transferability on downstream datasets and in-domain image-text retrieval tasks, compared to de-centralized pre-trained models. However, there are still performance gaps in the retrieval tasks compared to the centralized pre-trained model. That might because each batch of data in centralized pre-training scene has higher diversity, which encourages the contrastive-based model to capture more robust alignment.

Table 6: Downstream task performance on different multi-modal pre-training backbone methods.

Strategy	**Backbone**	**RSNA (cls.)**		**Covid (cls.)**		**RSNA (seg.)**		**In-domain Image-Text Retrieval**			
		1%	10%	1%	10%	1%	10%	**Rec.@1**	**Rec.@5**	**Wst.@1**	**Wst.@5**
Decentralized	ConVIRT	81.5	82.3	76.5	85.6	64.6	70.7	15.5	51.1	13.6	46.0
FedAvg	ConVIRT	83.1	83.3	78.0	88.5	69.6	71.5	28.8	72.1	25.3	66.7
Centralized	ConVIRT	83.4	84.6	82.5	92.0	72.6	76.4	41.5	84.2	38.6	80.0
Decentralized	GLoRIA	82.3	82.9	77.9	86.8	71.1	72.1	17.2	52.5	15.2	48.7
FedAvg	GLoRIA	83.2	83.3	77.5	89.0	71.4	72.4	29.9	73.8	27.8	69.5
Centralized	GLoRIA	84.0	84.7	82.2	91.8	73.6	73.7	41.7	84.0	39.0	80.5
Decentralized	MGCA	81.9	82.7	77.8	87.6	62.8	70.2	15.2	50.4	13.4	45.4
FedAvg	MGCA	82.6	83.5	75.8	88.2	70.1	71.4	29.3	73.7	26.8	70.4
Centralized	MGCA	84.0	84.5	79.5	89.5	70.7	72.5	39.9	83.5	36.9	80.3

C Detailed Experiment Results

Here we provide detailed results of ablation studies shown in Fig. 4 in the main text.

Table 7: Ablation studies on the number of clients.

Num. of Client	**RSNA (cls.)**		**Covid (cls.)**		**RSNA (seg.)**		**In-domain Image-Text Retrieval**			
	1%	10%	1%	10%	1%	10%	**Rec.@1**	**Rec.@5**	**Wst.@1**	**Wst.@5**
n=2	82.1	83.2	78.4	88.5	61.8	71.0	23.1	62.9	19.2	57.8
n=5	**83.2**	**83.7**	**81.0**	**90.3**	**71.7**	**74.1**	**30.2**	**73.2**	**27.0**	**68.9**

Table 8: Ablation studies on uncertainty radius.

Uncertainty Radius	**RSNA (cls.)**		**Covid (cls.)**		**RSNA (seg.)**		**In-domain Image-Text Retrieval**			
	1%	10%	1%	10%	1%	10%	**Rec.@1**	**Rec.@5**	**Wst.@1**	**Wst.@5**
ρ =0.01	82.7	83.2	79.6	89.1	71.0	72.8	**30.4**	**73.5**	26.6	68.4
ρ =0.1	83.2	83.7	81.0	90.3	71.7	**74.1**	30.2	73.2	**27.0**	**68.9**
ρ =1	**83.3**	**84.0**	**81.3**	**90.8**	**72.1**	74.1	28.9	72.5	26.2	67.8

Here we provided detailed results for our empirical study in Sec. 5.2.

Table 9: Ablation studies on global constraint degree.

Constraint Degree	**RSNA (cls.)**		**Covid (cls.)**		**RSNA (seg.)**		**In-domain Image-Text Retrieval**		
	1%	10%	1%	10%	1%	10%	**Rec.@1**	**Rec.@5**	**Disparity**
μ =1	82.8	83.4	79.8	89.6	70.5	72.8	29.1	72.5	3.2
μ =5	**83.2**	**83.7**	**81.0**	**90.3**	**71.7**	**74.1**	**30.2**	**73.2**	2.9
μ =10	82.6	83.2	80.2	90.2	71.3	72.9	29.6	72.8	**2.4**

Table 10: The comparison of retrieval acc. on each client denoted as $\{C_i\}_{i=1}^{5}$, of centralized, FedAvg, and averaged acc. of decentralized pre-trained models using the ConVIRT backbone. We report the local models of decentralized pre-training strategy as Decentralized$_i$.

Strategy	**Recall@1 (ACC)**						**Recall@5 (ACC)**					
	C1	**C2**	**C3**	**C4**	**C5**	**Avg.**	**C1**	**C2**	**C3**	**C4**	**C5**	**Avg.**
Centralized	43.6	38.6	40.1	43.1	41.9	44.4	86.6	80.0	82.6	85.0	86.8	84.2
FedAvg	30.4	25.3	26.9	28.8	32.7	28.8	76.4	66.7	69.8	73.8	73.9	72.1
Decentralized$_1$	17.7	14.7	15.4	18.2	14.1	16.0	57.0	49.6	51.3	54.6	55.1	53.5
Decentralized$_2$	15.3	11.6	13.9	15.0	13.8	13.9	54.8	41.9	46.2	47.9	45.6	47.3
Decentralized$_3$	17.4	13.1	14.1	15.1	15.4	15.0	50.4	44.2	46.4	49.7	52.5	48.6
Decentralized$_4$	16.5	14.3	14.1	15.4	17.4	15.5	57.0	45.7	47.5	52.4	57.6	52.0
Decentralized$_5$	21.7	14.3	14.2	15.6	19.5	17.1	57.9	48.4	50.2	52.4	61.7	54.1
Local.avg.	17.7	13.6	14.3	15.8	16.0	15.5	55.4	46.0	48.3	51.4	51.1	51.1

Table 11: The performance of the server model after 25 commu. turns and the averaged performance of corresponding local models after 25 and 26 commu. turns, on each client denoted as $\{C_i\}_{i=1}^{5}$. We utilize the ConVIRT as the backbone. We report the models after local update on client datasets of FedAvg pre-training strategy as Local$_i$.

Strategy	**com. turn**	**Recall@1 (ACC)**					**Recall@5 (ACC)**				
		C1	**C2**	**C3**	**C4**	**C5**	**C1**	**C2**	**C3**	**C4**	**C5**
FedAvg	25	30.4	25.3	26.9	28.8	32.7	76.4	66.7	69.8	73.8	73.9
Local$_0$	25	31.8	23.4	26.0	26.2	33.7	73.4	64.1	67.2	71.4	71.8
Local$_1$	25	27.7	22.5	23.8	25.1	27.7	73.4	63.1	64.6	69.2	68.6
Local$_2$	25	28.6	24.4	24.3	27.3	28.9	73.3	64.9	67.5	70.3	71.8
Local$_3$	25	30.6	22.6	23.9	25.4	26.4	72.6	64.0	66.3	68.9	67.6
Local$_4$	25	27.3	24.4	25.5	26.5	28.9	73.3	65.8	69.0	71.2	69.2
1-5 Avg.	25	29.2↓	23.5↓	24.7↓	26.1↓	29.1↓	73.2↓	64.4↓	66.9↓	70.2↓	69.8↓
Local$_0$	26	29.9	23.2	25.2	26.7	33.1	73.4	64.1	67.2	71.4	71.8
Local$_1$	26	28.7	23.1	24.0	25.9	27.4	74.5	63.4	64.1	69.7	67.9
Local$_2$	26	30.9	23.9	25.4	27.7	29.9	72.4	65.6	67.4	71.6	71.1
Local$_3$	26	30.1	22.7	23.4	24.5	27.1	73.1	63.5	65.3	67.5	67.3
Local$_4$	26	27.0	24.8	25.5	26.6	29.6	73.6	66.0	69.2	71.2	69.5
1-5 Avg.	26	29.3↓	23.5↓	24.7↓	26.3↓	29.4↓	73.3↓	64.4↓	66.3↓	70.1↓	69.4↓

Table 12: The accuracy of the test set of each client. We show the performance of FedAvg pre-trained baseline and its retrained models on different client datasets. We report the models after local update on client datasets of FedAvg pre-training strategy as Local_i.

position	**model**	**com.**	**Recall@1 (ACC)**						**Recall@5 (ACC)**					
			C0	**C1**	**C2**	**C3**	**C4**	**Avg.**	**C0**	**C1**	**C2**	**C3**	**C4**	**Avg.**
-	server	25	30.4	25.3	26.9	28.8	32.7	28.8	76.4	66.7	69.8	73.8	73.9	72.1
-	server	50	32.3	26.0	27.0	27.1	30.2	28.5↓	77.6	67.9	69.4	72.1	71.7	71.7↓
→ shallow	Local_0	25	30.4	25.0	25.3	28.4	28.6	27.5↓	73.8	67.2	68.4	72.0	73.0	70.9↓
→ shallow	Local_1	25	34.3	26.4	27.3	29.7	30.2	29.4↑	78.3	69.8	72.3	75.1	77.4	74.6↑
→ shallow	Local_2	25	33.7	26.4	27.3	29.7	30.2	29.4↑	77.3	67.4	70.7	74.2	70.5	72.0↑
→ shallow	Local_3	25	27.7	18.9	19.3	25.6	24.9	25.0↓	72.4	64.2	64.8	70.5	71.1	68.6↓
→ shallow	Local_4	25	26.2	18.9	19.3	22.7	22.7	22.0↓	69.3	56.8	58.9	63.3	64.5	62.6↓

Table 13: The accuracy of the model on each client. We show the acc. of centralized and FedAvg pre-trained baselines and de-centralized pre-trained models shown as Local_i retrained on the union of training splits of client datasets. We fine-tune shallow layers of the de-centralized pre-trained model with the union dataset.

strategy	**model**	**Recall@1 (ACC)**						**Recall@5 (ACC)**					
		C1	**C2**	**C3**	**C4**	**C5**	**Avg.**	**C1**	**C2**	**C3**	**C4**	**C5**	**Avg.**
Global	server	43.6	38.6	40.1	43.1	41.9	41.5	86.6	80.0	82.6	85.0	86.8	84.2
FedAvg	server	30.4	25.3	26.9	28.8	32.7	28.8	76.4	66.7	69.8	73.8	73.9	72.1
Decentralized	Local_1	28.7	22.6	23.5	22.3	24.6	24.4	70.9	60.8	63.3	62.4	63.4	64.2
Decentralized	Local_2	17.4	19.9	18.2	17.6	17.7	18.1	52.7	56.1	55.1	54.0	53.5	54.3
Decentralized	Local_3	20.9	20.7	26.0	21.0	22.3	22.2	58.9	58.1	65.3	58.2	59.0	59.9
Decentralized	Local_4	20.9	20.1	20.7	25.6	21.1	21.7	59.1	56.9	57.8	64.7	58.7	59.4
Decentralized	Local_5	21.8	19.5	22.0	20.9	31.5	23.2	60.7	57.0	59.8	60.2	74.1	62.4
Decentralized	Local_1	17.7	14.7	15.4	18.2	14.1	16.0	57.0	49.6	51.3	54.6	55.1	53.5
Decentralized	Local_2	15.3	11.6	13.9	15.0	13.8	13.9	54.8	41.9	46.2	47.9	45.6	47.3
Decentralized	Local_3	17.4	13.1	14.1	15.1	15.4	15.0	50.4	44.2	46.4	49.7	52.5	48.6
Decentralized	Local_4	16.5	14.3	14.1	15.4	17.4	15.5	57.0	45.7	47.5	52.4	57.6	52.0
Decentralized	Local_5	21.7	14.3	14.2	15.6	19.5	17.1	57.9	48.4	50.2	52.4	61.7	54.1

D Theoretical Analysis

D.1 Derivation of Proposition 1

To begin, we will establish the following lemma.

Lemma 1. *For a batch of samples z_a and z_b with batch size bz, and temperature parameter τ. Let $\overline{D}$ is the average L2 distance across the batch, $D_{\max}$ is the maximum L2 distance across the dataset. Suppose for all optimization batches there exist $L \geq 1, 0 \leq l \leq 1$, such that $l^2 \cdot D_{\max}^2 \leq \|z_{a,i} - z_{b,i}\|_2^2 \leq L^2 \cdot D_{\max}^2$, for $\forall z_{a,i}, z_{b,i}$ in the batch. Then the contrastive loss $L_{CL}(z_a, z_b)$ has the following upper bound:*

$$L_{CL}(z_a, z_b) \leq \log(bz) + \alpha \overline{D}^2,$$

where $\alpha = \frac{L^2(1-l^2)}{2\tau}$, is a client-specific constant.

Proof. The contrastive loss L_{CL} is given by

$$L_{\text{CL}}(z_a, z_b) = \frac{1}{\text{bz}} \sum_{i=1}^{\text{bz}} \left[-\log \frac{\exp(\text{sim}(z_{a,i}, z_{b,i})/\tau)}{\sum_{j=1}^{\text{bz}} \exp(\text{sim}(z_{a,i}, z_{b,j})/\tau)} \right],$$

where $\text{sim}(z_{a,i}, z_{b,i}) = \frac{z_{a,i} \cdot z_{b,i}}{\|z_{a,i}\| \|z_{b,i}\|}$. For normalized vectors $z_{a,i}$ and $z_{b,i}$, we have:

$$\|z_{a,i} - z_{b,i}\|_2^2 = 2 - 2 \cdot \text{sim}(z_{a,i}, z_{b,i}), \quad \Rightarrow \quad \text{sim}(z_{a,i}, z_{b,i}) = 1 - \frac{1}{2}\|z_{a,i} - z_{b,i}\|_2^2.$$

Substituting $\text{sim}(z_{a,i}, z_{b,i})$ into L_{CL}:

$$L_{\text{CL}}(z_a, z_b) = \frac{1}{\text{bz}} \sum_{i=1}^{\text{bz}} \left[-\log \frac{\exp\left((1 - \frac{1}{2}\|z_{a,i} - z_{b,i}\|_2^2)/\tau\right)}{\sum_{j=1}^{\text{bz}} \exp\left((1 - \frac{1}{2}\|z_{a,i} - z_{b,j}\|_2^2)/\tau\right)} \right].$$

Let $D_{\max}^2 = \max_i \|z_{a,i} - z_{b,i}\|_2^2$, we have:

$$L_{\text{CL}}(z_a, z_b) \leq \log(\text{bz}) + \frac{1}{2\tau} \left(D_{\max}^2 - \frac{1}{\text{bz}} \sum_{i=1}^{\text{bz}} \|z_{a,i} - z_{b,i}\|_2^2 \right).$$

Substituting the assumption $l^2 \cdot D_{\max}^2 \leq \|z_{a,i} - z_{b,i}\|_2^2 \leq L^2 \cdot D_{\max}^2$, we get:

$$L_{\text{CL}}(z_a, z_b) \leq \log(\text{bz}) + \frac{1}{2\tau} \left(D_{\max}^2 - \frac{1}{\text{bz}} \sum_{i=1}^{\text{bz}} l^2 \cdot D_{\max}^2 \right).$$

This simplifies further to:

$$L_{\text{CL}}(z_a, z_b) \leq \log(\text{bz}) + \frac{1}{2\tau} \cdot D_{\max}^2 \cdot (1 - l^2).$$

Now, since we have $D_{\max} \leq L \cdot \overline{D}$, where $\overline{D} = \frac{1}{\text{bz}} \sum_{i=1}^{\text{bz}} \|z_{a,i} - z_{b,i}\|_2^2$. Substituting this into the inequality above:

$$L_{\text{CL}}(z_a, z_b) \leq \log(\text{bz}) + \frac{1}{2\tau} \cdot (l \cdot \overline{D})^2 \cdot (1 - l^2).$$

Replacing $\alpha = \frac{L^2(1-l^2)}{2\tau}$, we have:

$$L_{\text{CL}}(z_a, z_b) \leq \log(\text{bz}) + \alpha \overline{D}^2.$$

□ □

Using this lemma, we will complete the proof of Proposition 1. In this paper, without loss of generalizability, we assume $\mathcal{R}_T(\cdot)$ to be the contrastive loss.

Proof. We begin by expressing the generalization error $\mathcal{R}_T(\hat{f})$ on the target domain $\mathcal{D}_\mathcal{T}$ as the expected contrastive loss:

$$\mathcal{R}_T(\hat{f}) = \mathbb{E}_{(x,y)\sim\mathcal{D}_\mathcal{T}}\left[L_{\text{CL}}(\hat{f}_\psi(x), \hat{f}_\phi(y))\right], \tag{3}$$

where L_{CL} is the contrastive loss defined as:

$$L_{\text{CL}}(z_a, z_b) = \frac{1}{\text{bz}}\sum_{i=1}^{\text{bz}}\left[-\log\frac{\exp\left(\text{sim}(z_{a,i}, z_{b,i})/\tau\right)}{\sum_{j=1}^{\text{bz}}\exp\left(\text{sim}(z_{a,i}, z_{b,j})/\tau\right)}\right], \tag{4}$$

with $\text{sim}(z_{a,i}, z_{b,i}) = \frac{z_{a,i}^\top z_{b,i}}{\|z_{a,i}\|\|z_{b,i}\|}$ and bz being the batch size.

By Lemma 1, we have an upper bound on the contrastive loss:

$$L_{\text{CL}}(z_a, z_b) \leq \log(\text{bz}) + \alpha\overline{D}^2, \tag{5}$$

where $\overline{D}^2$ is the average squared Euclidean distance between $z_{a,i}$ and $z_{b,i}$:

$$\overline{D}^2 = \frac{1}{\text{bz}}\sum_{i=1}^{\text{bz}}\|z_{a,i} - z_{b,i}\|_2^2. \tag{6}$$

Applying this to our generalization error:

$$\mathcal{R}_T(\hat{f}) \leq \log(\text{bz}) + \alpha\mathbb{E}_{(x,y)\sim\mathcal{D}_\mathcal{T}}\left[\|\hat{f}_\psi(x) - \hat{f}_\phi(y)\|_2^2\right]. \tag{7}$$

Then, we have:

$$\begin{aligned}
|\hat{f}_\psi(x) - \hat{f}_\phi(y)|_2^2 \leq &\Big(|\hat{f}_\psi(x) - \hat{f}_{\psi_i}(x)|_2 + |\hat{f}_{\psi_i}(x) - f_{\psi_i}(x)|_2 \\
&+|f_{\psi_i}(x) - f_{\phi_i}(y)|_2 + |f_{\phi_i}(y) - \hat{f}_{\phi_i}(y)|_2 + |\hat{f}_{\phi_i}(y) - \hat{f}_\phi(y)|_2\Big)^2 \\
\leq 5&\Big(|\hat{f}_\psi(x) - \hat{f}_{\psi_i}(x)|_2^2 + |\hat{f}_{\psi_i}(x) - f_{\psi_i}(x)|_2^2 \\
&+|f_{\psi_i}(x) - f_{\phi_i}(y)|_2^2 + |f_{\phi_i}(y) - \hat{f}_{\phi_i}(y)|_2^2 + |\hat{f}_{\phi_i}(y) - \hat{f}_\phi(y)|_2^2\Big),
\end{aligned} \tag{8}$$

where the last inequality follows from the fact that for any real numbers $a_1, \ldots, a_n$,

$$\left(\sum_{j=1}^n a_j\right)^2 \leq n\sum_{j=1}^n a_j^2.$$

Define the error terms:

$$\begin{aligned}
\epsilon_{\hat{f}_\psi, \hat{f}_{\psi_i}}(x) &= |\hat{f}_\psi(x) - \hat{f}_{\psi_i}(x)|_2^2, \\
\epsilon_{\hat{f}_{\psi_i}, f_{\psi_i}}(x) &= |\hat{f}_{\psi_i}(x) - f_{\psi_i}(x)|_2^2, \\
\epsilon_{\hat{f}_\phi, \hat{f}_{\phi_i}}(y) &= |\hat{f}_\phi(y) - \hat{f}_{\phi_i}(y)|_2^2, \\
\epsilon_{\hat{f}_{\phi_i}, f_{\phi_i}}(y) &= |\hat{f}_{\phi_i}(y) - f_{\phi_i}(y)|_2^2, \\
C_i(x, y) &= |f_{\psi_i}(x) - f_{\phi_i}(y)|_2^2.
\end{aligned}$$

Then inequality (8) becomes:

$$|\hat{f}_\psi(x) - \hat{f}_\phi(y)|_2^2 \leq 5\left(\epsilon_{\hat{f}_\psi,\hat{f}_{\psi_i}}(x) + \epsilon_{\hat{f}_{\psi_i},f_{\psi_i}}(x) + C_i(x,y) + \epsilon_{\hat{f}_{\phi_i},f_{\phi_i}}(y) + \epsilon_{\hat{f}_\phi,\hat{f}_{\phi_i}}(y)\right). \tag{9}$$

Taking expectation over $(x,y) \sim \mathcal{D}_\mathcal{T}$ and using $\mathcal{D}_\mathcal{T} = \sum_{i=1}^N w_i \mathcal{D}_i$, we have:

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\mathcal{D}_\mathcal{T}}\left[|\hat{f}_\psi(x) - \hat{f}_\phi(y)|_2^2\right] &= \sum_{i=1}^N w_i\, \mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[|\hat{f}_\psi(x) - \hat{f}_\phi(y)|_2^2\right] \\ &\leq 5\sum_{i=1}^N w_i\, \mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\epsilon_{\hat{f}_\psi,\hat{f}_{\psi_i}}(x) + \epsilon_{\hat{f}_{\psi_i},f_{\psi_i}}(x)\right. \\ &\quad \left. + C_i(x,y) + \epsilon_{\hat{f}_{\phi_i},f_{\phi_i}}(y) + \epsilon_{\hat{f}_\phi,\hat{f}_{\phi_i}}(y)\right]. \end{aligned} \tag{10}$$

Define $C_i = \mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[C_i(x,y)\right]$. Then,

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\mathcal{D}_\mathcal{T}}\left[|\hat{f}_\psi(x) - \hat{f}_\phi(y)|_2^2\right] \leq 5\sum_{i=1}^N w_i \Big(&\mathbb{E}_{x\sim\mathcal{D}_i}\left[\epsilon_{\hat{f}_\psi,\hat{f}_{\psi_i}}(x) + \epsilon_{\hat{f}_{\psi_i},f_{\psi_i}}(x)\right] \\ &+ \mathbb{E}_{y\sim\mathcal{D}_i}\left[\epsilon_{\hat{f}_{\phi_i},f_{\phi_i}}(y) + \epsilon_{\hat{f}_\phi,\hat{f}_{\phi_i}}(y)\right] + C_i\Big). \end{aligned} \tag{11}$$

Substituting (11) into the generalization error bound, we obtain:

$$\begin{aligned}\mathcal{R}_T(\hat{f}) \leq \log(\text{bz}) + \alpha\sum_{i=1}^N w_i \Big(&5\mathbb{E}_{x\sim\mathcal{D}_i}\left[\epsilon_{\hat{f}_\psi,\hat{f}_{\psi_i}}(x) + \epsilon_{\hat{f}_{\psi_i},f_{\psi_i}}(x)\right] \\ &+ 5\mathbb{E}_{y\sim\mathcal{D}_i}\left[\epsilon_{\hat{f}_{\phi_i},f_{\phi_i}}(y) + \epsilon_{\hat{f}_\phi,\hat{f}_{\phi_i}}(y)\right] + 5C_i\Big). \end{aligned} \tag{12}$$

Letting $\alpha_i = 5\alpha$, we have:

$$\mathcal{R}_T(\hat{f}) \leq \sum_{i=1}^N w_i\alpha_i \left(\mathbb{E}_{x\sim\mathcal{D}_i}\left[\epsilon_{\hat{f}_\psi,\hat{f}_{\psi_i}}(x) + \epsilon_{\hat{f}_{\psi_i},f_{\psi_i}}(x)\right] + \mathbb{E}_{y\sim\mathcal{D}_i}\left[\epsilon_{\hat{f}_{\phi_i},f_{\phi_i}}(y) + \epsilon_{\hat{f}_\phi,\hat{f}_{\phi_i}}(y)\right] + C_i\right) + \log(\text{bz}). \tag{13}$$

Since $\log(\text{bz})$ is independent of $\hat{f}$, it can be considered a constant. Thus, we can express the generalization error as:

$$\mathcal{R}_T(\hat{f}) \leq \sum_{i=1}^N w_i\alpha_i \left(\epsilon_{\hat{f}_\psi,\hat{f}_{\psi_i}} + \epsilon_{\hat{f}_{\psi_i},f_{\psi_i}} + \epsilon_{\hat{f}_\phi,\hat{f}_{\phi_i}} + \epsilon_{\hat{f}_{\phi_i},f_{\phi_i}} + C_i\right).$$

This completes the proof of Proposition 1. □