FEDMAGE: ENHANCING FEDERATED SELF-SUPERVISED LEARNING UNDER HETEROGENEOUS DATA WITH GENERATIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated self-supervised learning enables representation learning from decentralized unlabeled data, but non-IID client distributions can induce inconsistent local representations and weak global coordination. Existing methods mitigate this issue by communicating compact representation summaries, such as prototypes or centroids, yet such compression retains only partial information about client data. Generative learning offers a promising alternative through richer sample-level surrogates, but its representation-level effect at both clients and the server remains unclear. In this paper, we theoretically characterize generative enhancement in federated masked image modeling and show that local generative modeling strengthens the recovery of globally relevant semantic directions under heterogeneous data. We further show that semantically consistent generated data improve global representability beyond conventional parameter aggregation. Building on these insights, we propose FedMAGE, a theory-guided framework that combines masked reconstruction and variational generation through a shared client encoder, while using quality-aware client generators to refine the aggregated global model. Extensive experiments show consistent gains over representative baselines across varying levels of data heterogeneity, while ablation studies confirm the contribution of the main components of FedMAGE.

1 INTRODUCTION

Federated learning (FL) enables distributed clients to collaboratively learn a global model without directly sharing raw data, making it well suited to privacy-sensitive applications (McMahan et al., 2017; Zhang et al., 2021). In practice, however, much client-side data are unlabeled and expensive to annotate (Verbraeken et al., 2020). Self-supervised learning (SSL) provides a natural solution by learning transferable representations directly from unlabeled observations (Liu et al., 2021; Gui et al., 2024), motivating growing interest in federated self-supervised learning (FedSSL) (Zhuang et al., 2021a; Wang et al., 2022; Chen et al., 2026). Existing FedSSL methods extend representative objectives such as contrastive learning (Chen et al., 2020; Chen & He, 2021) and masked reconstruction (He et al., 2022; Bao et al., 2021) to distributed settings.

A central challenge in FedSSL is data heterogeneity, where non-IID client distributions can drive local models toward different representation structures and weaken global aggregation (Zhu et al., 2021). Existing studies therefore improve cross-client consistency by communicating compact representation-level information, such as prototypes (Tan et al., 2022), cluster centroids (Lubana et al., 2022), or embedding statistics (Chen et al., 2026). These summaries can capture useful cross-client semantic structure and provide global guidance, but they inevitably compress many local observations into a small set of statistics or representative vectors (Fu et al., 2025). As a result, the server still receives only a partial view of sample-level variation and fine-grained structure, leaving an information bottleneck in global representation coordination.

Generative learning offers a natural way to reduce this bottleneck. Locally learned generators can synthesize observations that reflect characteristics of client distributions while keeping the source data local (Goodfellow et al., 2020; Ho et al., 2020; Kingma & Welling, 2014). Such synthetic samples may retain richer instance-specific information than compact statistics and provide a more

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

informative basis for global coordination. However, whether this additional information actually improves federated representation learning under heterogeneous data remains unclear. In particular, it is unknown whether generated samples can counteract representation drift and refine the global representation toward the true semantic structure distributed across clients.

Meanwhile, the usefulness of synthetic data depends critically on the quality of the generator that produces them (He et al., 2025). In the more general setting where a pretrained generative model is unavailable, a practical federated solution must learn generative capability alongside self-supervised representation learning on local client data. This joint training changes the local objective and introduces an additional learning signal, raising the question of how generation actually affects representation learning under non-IID data. Together with the unresolved role of synthetic data in refining the aggregated global model, these two gaps motivate the central question of our study:

Can generative learning help FedSSL learn stronger representations under heterogeneous data and enable more effective global coordination?

In this paper, we address this question by establishing a theoretical characterization of generative enhancement in federated SSL. Our analysis shows that augmenting local masked image modeling (MIM) with generative learning provides stronger guarantees for recovering globally relevant semantic directions under data heterogeneity. We further show that, when generated data preserve distributed semantic statistics, server-side synthetic refinement yields stronger global representability guarantees than parameter aggregation alone. Guided by these findings, we develop FedMAGE, a federated masked representation learning framework that combines local masked-generative learning with quality-guided global refinement. Each client jointly performs masked reconstruction and variational generation, evaluates the semantic and structural quality of its generator, and uploads the resulting scores with its model update. The server then aggregates client models, selects reliable generators, and uses their synthetic samples to refine the aggregated global model. Extensive experiments across multiple datasets and non-IID federated settings show consistent gains over representative baselines, while robustness and ablation studies support the theoretical findings and key design choices.

Overall, our main contributions are summarized as follows:

- To the best of our knowledge, we provide the first representation-level analysis of generative enhancement in federated SSL, showing that local generative learning strengthens semantic recoverability under heterogeneous data and that semantically consistent synthetic information improves global representability beyond standard parameter aggregation.
- We propose FedMAGE, a theory-guided framework that combines masked representation learning with lightweight local generative modeling and quality-aware server-side refinement, enabling richer cross-client coordination without exchanging raw observations.
- We conduct extensive experiments across diverse baselines, datasets, and levels of data heterogeneity, demonstrating the effectiveness and robustness of FedMAGE and clarifying the contributions of its key design components.

2 Related Work

2.1 Federated Self-Supervised Learning

FedSSL has become an important paradigm for learning representations from decentralized unlabeled data, with a major challenge being the degradation caused by client heterogeneity (McMahan et al., 2017; Zhu et al., 2021). Existing works study this issue from both algorithmic and theoretical perspectives. Algorithmically, recent methods improve cross-client consistency by exchanging compact representation summaries, such as prototypes or cluster-level statistics (Tan et al., 2022; Lubana et al., 2022; Wang et al., 2022). Theoretically, prior work has analyzed how heterogeneity affects distributed representation learning (Chen et al., 2026) or established convergence guarantees for specific FedSSL algorithms (Lubana et al., 2022; Liao et al., 2024). However, these analyses mainly consider conventional SSL objectives and training processes (Xie et al., 2022; He et al., 2022), leaving the representation-level effect of incorporating a complementary learning paradigm unexplored. Our work extends this direction by analyzing generative enhancement in federated MIM and using the resulting insights to guide a new coordination mechanism.

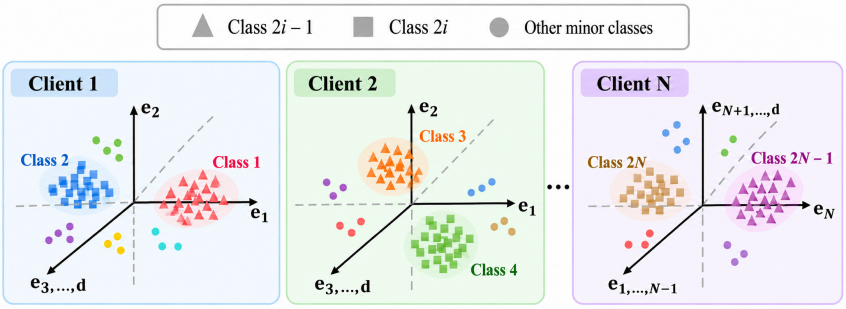

Figure 1: **Illustration of the heterogeneous data modeling used in analysis.**

2.2 Generative Learning in Federated Settings

Generative modeling has progressed from variational autoencoders (VAEs) (Kingma & Welling, 2014) and generative adversarial networks (GANs) (Goodfellow et al., 2020) to diffusion models (Ho et al., 2020), substantially improving the ability to model complex data distributions and synthesize realistic samples. It has also been extended to federated settings, since vast amounts of distributed data can support effective learning (Liu et al., 2024). Existing work broadly follows two directions. One treats generation as the primary objective and studies how to train high-quality generators from decentralized and heterogeneous data (Stanley Jothiraj & Mashhadi, 2024; Xiong et al., 2023b). The other uses generation as an auxiliary mechanism to improve FL, including enriching cross-domain data coverage (Zhang et al., 2025), compensating for missing modalities (Xiong et al., 2023a), and strengthening augmentation or auxiliary supervision in self-supervised learning (Belagali et al., 2026; Li et al., 2023). These studies primarily emphasize the task-specific utility of generation, while our work goes beyond empirical use of generative components, providing principled clarification for when and how generative enhancement can benefit FedSSL.

3 Theoretical Analysis of Generative Enhancement

In this section, we provide a formal analysis of generative enhancement in FedSSL. We first introduce the heterogeneous data model and training process used throughout the analysis. We then address the central research question by analyzing how local generative objectives affect representation learning and whether server-side synthetic refinement further improves global representability.

3.1 Problem Setup

Client-Side Heterogeneity and Local Learning. We study a federated system with N clients, where client $i \in [N]$ holds an unlabeled dataset D_i drawn from the local distribution $\mathcal{D}_i$. Assuming equal local sample sizes, the global distribution is $\mathcal{D} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{D}_i$. Following prior theoretical studies of distributed SSL under non-IID data (Wang et al., 2022; Liu et al., 2022; Chen et al., 2026), we adopt a simplified but rigorous latent-class model with $2N$ global classes to capture client-specific semantic heterogeneity. As illustrated in Figure 1, client i is dominated by classes $2i-1$ and $2i$, which contain the same number of samples and grow polynomially with the ambient dimension d, whereas all remaining classes together contain only $O(d^\alpha)$ samples for some $\alpha \in (0,1)$. Let $e_1, \ldots, e_d$ denote the standard basis of $\mathbb{R}^d$, with d sufficiently large. Samples from the two dominant classes are constructed as

$$x^{(2i-1)} = e_i - \sum_{k\neq i} q^{(2i-1,k)} \tau e_k + \mu \xi^{(2i-1)}, \quad x^{(2i)} = -e_i - \sum_{k\neq i} q^{(2i,k)} \tau e_k + \mu \xi^{(2i)}, \tag{1}$$

where $q^{(\cdot,\cdot)} \sim \text{Bernoulli}(1/2)$ are independently sampled, $\tau > 0$ controls cross-client semantic interference, $\mu > 0$ scales sample-level noise, and $\xi^{(\cdot)} \sim \mathcal{N}(0, I_d)$ captures random variation. This construction makes each local distribution primarily governed by client-specific semantic directions while collectively preserving coverage of the global semantic space.

Under this setting, each client trains a shared linear encoder $W \in \mathbb{R}^{c\times d}$ and decoder $V \in \mathbb{R}^{d\times c}$ for masked reconstruction and generation. Given a binary mask $m \in \{0,1\}^d$ whose entries are independently masked with probability p_m, we define $x^{\text{vis}} = x \odot m$ and $x^{\text{mask}} = x \odot (1-m)$. The local MIM objective is

$$\mathcal{L}^{(i)}_{\text{MIM}}(W,V) = \mathbb{E}_{x\sim\mathcal{D}_i,m}\left[\left\|VWx^{\text{vis}} - x^{\text{mask}}\right\|_2^2\right]. \tag{2}$$

To introduce generative learning within the same representation space, we further instantiate the encoder-decoder pair as a linear-Gaussian VAE (Kingma & Welling, 2014) with $q_W(z \mid x) = \mathcal{N}(Wx, \sigma^2 I_c)$ and prior $p(z) = \mathcal{N}(0, I_c)$. Its objective $\mathcal{L}^{(i)}_{\text{VAE}}$ combines input reconstruction $\mathbb{E}\|Vz - x\|_2^2$ with KL regularization $D_{\text{KL}}(q_W(z \mid x)\|p(z))$. The resulting local objective is

$$\mathcal{L}^{(i)}_{\text{loc}}(W,V) = \mathcal{L}^{(i)}_{\text{MIM}}(W,V) + \lambda\mathcal{L}^{(i)}_{\text{VAE}}(W,V), \tag{3}$$

where $\lambda > 0$ controls the generative loss and $\beta > 0$ weights the KL term in $\mathcal{L}^{(i)}_{\text{VAE}}$. For analytical simplicity, we set $\lambda = \beta = 1$ to remove unnecessary scaling factors from the proof and denote the locally optimized parameters by (W_i, V_i).

Server-Side Aggregation and Global Refinement. After local optimization, the server follows standard FedAvg (McMahan et al., 2017) to aggregate client models. Since all clients have equal-sized local datasets, uniform aggregation gives $\bar{W} = \frac{1}{N}\sum_{i=1}^N W_i$ and $\bar{V} = \frac{1}{N}\sum_{i=1}^N V_i$. The locally learned generative decoders further allow the server to synthesize client-specific data. For client i, we draw $z_i \sim \mathcal{N}(0, I_c)$ and generate $\tilde{x}_i = V_i z_i$, inducing a synthetic distribution $\widetilde{\mathcal{D}}_i$. Sampling equally from all client generators yields $\widetilde{\mathcal{D}} = \frac{1}{N}\sum_{i=1}^N \widetilde{\mathcal{D}}_i$. Starting from the aggregated model $(\bar{W}, \bar{V})$, the server performs masked reconstruction on $\widetilde{\mathcal{D}}$ using the same masking mechanism as local MIM:

$$\mathcal{L}_{\text{srv}}(W,V) = \mathbb{E}_{\tilde{x}\sim\widetilde{\mathcal{D}},m}\left[\|VW(\tilde{x}\odot m) - \tilde{x}\odot(1-m)\|_2^2\right]. \tag{4}$$

Thus, one communication round follows local optimization, parameter aggregation, and server-side synthetic refinement. For theoretical tractability, rather than explicitly tracking intermediate parameter iterates, we characterize the supervision introduced by local learning and synthetic refinement through the one-round surrogate objective

$$\mathcal{L}_{\text{Global}}(W,V) = \frac{1}{N}\sum_{i=1}^N \mathcal{L}^{(i)}_{\text{loc}}(W,V) + \mathcal{L}_{\text{srv}}(W,V). \tag{5}$$

This surrogate enables direct comparison between conventional federated training and training augmented with server-side synthetic refinement. The complete formulation and technical assumptions are provided in Appendix A.6.1 and A.6.2.

3.2 Generative Learning Improves Local and Global Representability

We next analyze the representation-level effect of generative learning. Inspired by prior theoretical works on SSL (Wang et al., 2022; Liu et al., 2022; Chen et al., 2026), we use the representability vector (RV) to quantify how the learned representation changes under different objectives. For an encoder W, we define $r(W) = [r_1(W), \ldots, r_d(W)]$, where $r_k(W) = \|\Pi_{\text{row}(W)}(e_k)\|_2^2$ measures how strongly semantic direction e_k is captured by the encoder row space. A larger $r_k(W)$ indicates stronger semantic recovery. This measure is particularly informative under non-IID data, where client-specific dominant directions can bias the learned representation and suppress globally relevant semantics. We therefore use RV to compare conventional federated MIM, local masked-generative learning, and the complete framework with server-side synthetic refinement.

Theorem 3.1. (Enhanced Representability through Generative Learning). *Consider a federated system with $N = \Theta(d^{1/20})$ clients following the heterogeneous data setup, with $\tau = d^{1/5}$ and $\mu = d^{-1/5}$. Let $\delta_d = \max\left\{\delta_d^{\text{MIM}}, \delta_d^{\text{Gen}}, \delta_d^{\text{Global}}\right\}$ be a common upper bound on the corresponding population-to-empirical representation perturbations. For any fixed $p_m \in (0,1)$ and sufficiently large d, the following results hold relative to the representation learned by federated MIM.*

1. *Let* $r_k^{\mathrm{Gen}} = \|\Pi_{\mathrm{row}(W_{\mathrm{Gen}})} e_k\|_2^2$ *denote the representability after local generative learning but before server-side refinement. Then, for every* $e_k \in \mathcal{S}$*,*

$$r_k^{\mathrm{Gen}} \geq 1 - \frac{4\delta_d^2}{\left[C_{\mathrm{Gen}}(p_m) d^{2/5} + o(d^{2/5})\right]^2} > 1 - \frac{4\delta_d^2}{\left[C_{\mathrm{MIM}}(p_m) d^{2/5} + o(d^{2/5})\right]^2}, \quad (6)$$

where $C_{\mathrm{Gen}}(p_m) = \frac{[1-p_m(1-p_m)]^2}{4(2-p_m^2)}$ *and* $C_{\mathrm{MIM}}(p_m) = \frac{p_m^2(1-p_m)}{4(1+p_m)}$.

2. *When the aggregated synthetic distribution consistently preserves the global semantic statistics, let* $r_k^{\mathrm{Global}} = \|\Pi_{\mathrm{row}(W_{\mathrm{Global}})} e_k\|_2^2$ *denote the representability after server-side generative refinement. Then, for every* $e_k \in \mathcal{S}$*,*

$$r_k^{\mathrm{Global}} \geq 1 - \frac{4\delta_d^2}{\left[C_{\mathrm{Global}}(p_m) d^{2/5} + o(d^{2/5})\right]^2} > 1 - \frac{4\delta_d^2}{\left[C_{\mathrm{MIM}}(p_m) d^{2/5} + o(d^{2/5})\right]^2}, \quad (7)$$

where $C_{\mathrm{Global}}(p_m) = \frac{[1-2p_m(1-p_m)]^2}{4(3-2p_m^2)}$.

Proof Sketch. We establish the representability bound for the generative-enhanced framework by rewriting the masked and generative objectives in terms of second-order statistics, eliminating the optimal decoder, and characterizing the dominant semantic eigenspace of the resulting population representation operator. We then derive the FedMIM baseline by removing the complete-data VAE statistics and server-side synthetic terms, yielding $\Delta_{\mathrm{FedMIM}} \geq C_{\mathrm{MIM}}(p_m) d^{2/5} + o(d^{2/5})$. Using a common perturbation scale, local generative learning gives $\Delta_{\mathrm{Gen}} \geq C_{\mathrm{Gen}}(p_m) d^{2/5} + o(d^{2/5})$, while semantically consistent server-side refinement gives $\Delta_{\mathrm{Global}} \geq C_{\mathrm{Global}}(p_m) d^{2/5} + o(d^{2/5})$. Since $C_{\mathrm{Gen}}(p_m) > C_{\mathrm{MIM}}(p_m)$ and $C_{\mathrm{Global}}(p_m) > C_{\mathrm{MIM}}(p_m)$ for all $p_m \in (0,1)$, standard matrix concentration (Tropp, 2012) and the Davis–Kahan eigenspace perturbation theorem (Davis & Kahan, 1970) convert these larger semantic separations into the tighter representability guarantees in Theorem 3.1. The full derivation is provided in Appendix A.6.3.

Theorem 3.1 provides a unified representation-level answer to the research question motivating our study. First, the tighter bound associated with $C_{\mathrm{Gen}}(p_m)$ shows that introducing a local generative objective helps preserve globally relevant semantic directions that can otherwise be weakened by client-specific data heterogeneity. This suggests that generative learning does more than provide an auxiliary reconstruction signal. By incorporating complete-data statistics, it improves the robustness of local representation learning to semantic imbalance. Second, the improvement associated with $C_{\mathrm{Global}}(p_m)$ shows that server-side synthetic refinement can also strengthen the global representation beyond parameter aggregation, provided that the generated data preserve the distributed semantic statistics with sufficient fidelity. This highlights that the benefit of synthetic data depends not only on their availability but also on their semantic quality. These two findings suggest that generative modeling is naturally suited to federated SSL: it can improve local semantic recovery while providing reliable additional information for global refinement, thereby supporting more robust representation learning under heterogeneous data.

4 FedMAGE

Building on the theoretical findings above, we design FedMAGE with two complementary stages of generative enhancement. As illustrated in Figure 2, the first performs local masked-generative learning by incorporating generative modeling into client-side self-supervised pretraining. The second refines the aggregated global model using quality-guided synthetic data. We detail both stages below.

4.1 Local Masked-Generative Representation Learning

FedMAGE extends MIM with a client-specific variational generation branch. At communication round t, each selected client i initializes its local model from the current global parameters, consisting of a shared encoder E_i, reconstruction decoder R_i, and image generator $\mathcal{G}_i$. For an unlabeled image $x \sim \mathcal{D}_i$, we randomly mask patch indices Ω and encode the visible patches. The reconstruction decoder predicts the removed content as

$$\mathcal{L}_{\mathrm{MIM}}^{(i)} = \frac{1}{|\Omega|} \sum_{j \in \Omega} \|\hat{x}_{i,j} - x_j\|_2^2, \qquad \hat{x}_i = R_i\big(E_i(x^{\mathrm{vis}})\big). \quad (8)$$

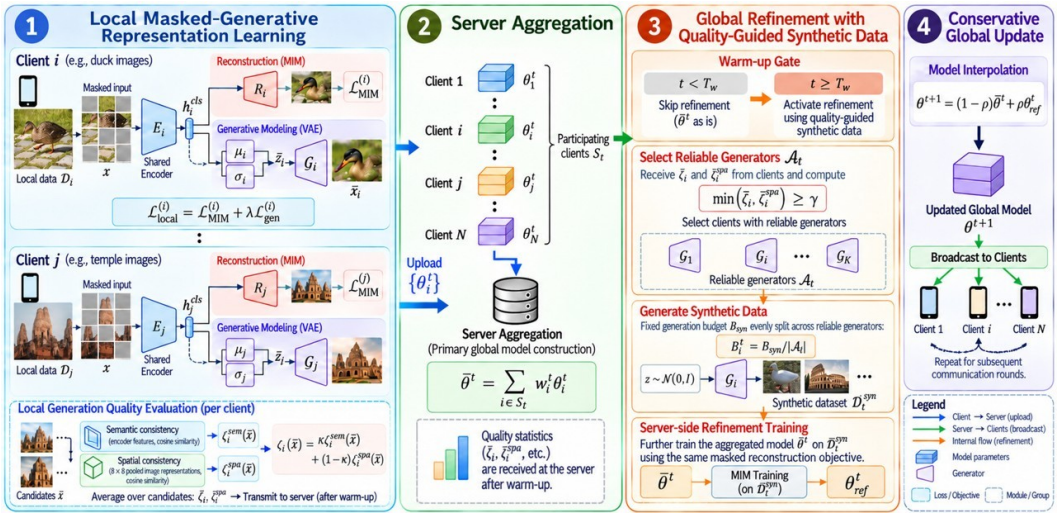

Figure 2: **Overview of FedMAGE.** The figure shows one round of the training process.

Then, FedMAGE models the local image distribution from the resulting `[CLS]` token h_i^{cls} obtained from the visible patches. Two lightweight projections parameterize a Gaussian posterior, with $\nu_i = f_{\nu,i}(h_i^{\text{cls}})$ and $\ell_i = f_{\ell,i}(h_i^{\text{cls}})$ denoting its mean and log-variance. Thus, $Q_i(z \mid x^{\text{vis}}) = \mathcal{N}(\nu_i, \text{diag}(\exp(\ell_i)))$, and the latent variable is sampled by reparameterization, $z_i = \nu_i + \exp\left(\frac{\ell_i}{2}\right) \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. The generator then reconstructs a complete image, $\tilde{x}_i = \mathcal{G}_i(z_i)$. To model the local distribution while enabling prior-based sampling, we optimize

$$\mathcal{L}_{\text{gen}}^{(i)} = \|\tilde{x}_i - x\|_2^2 + \beta D_{\text{KL}}\left(Q_i(z \mid x^{\text{vis}}) \,\|\, \mathcal{N}(0, I)\right). \tag{9}$$

The local objective is

$$\mathcal{L}_{\text{local}}^{(i)} = \mathcal{L}_{\text{MIM}}^{(i)} + \lambda \mathcal{L}_{\text{gen}}^{(i)}, \tag{10}$$

where λ controls generative learning and β weights KL regularization. The reconstruction decoder and generator form separate branches over the shared encoder. When enabled, the generative gradient is stopped at h_i^{cls} to prevent skewed local data from excessively shifting the transferable encoder.

Since the server later uses client generators to construct synthetic data, FedMAGE evaluates their quality during local optimization. For each real local sample x, the corresponding generated reconstruction $\tilde{x}$ is compared with its source image to assess whether the generator preserves its semantic content and spatial structure. Specifically, semantic and spatial consistency are measured by

$$\zeta_i^{\text{sem}}(\tilde{x}, x) = \cos(\phi_i^{\text{sem}}(\tilde{x}), \phi_i^{\text{sem}}(x)), \quad \zeta_i^{\text{spa}}(\tilde{x}, x) = \cos(\phi^{\text{spa}}(\tilde{x}), \phi^{\text{spa}}(x)), \tag{11}$$

where $\phi_i^{\text{sem}}(\cdot)$ denotes the encoder-derived semantic feature and $\phi^{\text{spa}}(\cdot)$ the flattened representation after 8×8 adaptive average pooling. We combine the two scores as

$$\zeta_i(\tilde{x}, x) = \kappa\, \zeta_i^{\text{sem}}(\tilde{x}, x) + (1 - \kappa)\, \zeta_i^{\text{spa}}(\tilde{x}, x), \tag{12}$$

where $\kappa \in [0, 1]$ balances semantic and spatial consistency. The combined and spatial scores are averaged over local evaluation samples and transmitted with the model once synthetic refinement is activated. These statistics provide the reliability information used for server-side generator selection.

4.2 Global Refinement with Quality-Guided Synthetic Data

After local training, the server aggregates the returned client models. Let $\mathcal{S}_t$ denote the participating clients at round t and θ_i^t the parameters returned by client i. The aggregated parameters are

$$\bar{\theta}^t = \sum_{i \in \mathcal{S}_t} w_i^t \theta_i^t, \qquad \sum_{i \in \mathcal{S}_t} w_i^t = 1, \tag{13}$$

where w_i^t is the normalized aggregation weight. FedMAGE retains parameter aggregation and uses $\bar{\theta}^t$ to initialize a subsequent generative refinement stage.

Refinement starts after a warm-up of T_{w} rounds, allowing client generators to first learn meaningful local distributions. Thus, for $t < T_{\mathrm{w}}$, training uses parameter aggregation alone. Once $t \geq T_{\mathrm{w}}$, participating clients additionally provide their generation-quality statistics. Let $\bar{\zeta}_i$ and $\bar{\zeta}_i^{\mathrm{spa}}$ denote the averaged combined and spatial confidence scores of $\mathcal{G}_i$. The server retains $\mathcal{G}_i$ if

$$\min\left(\bar{\zeta}_i, \bar{\zeta}_i^{\mathrm{spa}}\right) \geq \gamma, \tag{14}$$

where $\gamma \in [0, 1]$ is the quality threshold. This forms the reliable generator set $\mathcal{A}_t \subseteq \{\mathcal{G}_i\}_{i \in \mathcal{S}_t}$. Given a fixed generation budget B_{syn}, each retained generator receives

$$B_i^t = \frac{B_{\mathrm{syn}}}{|\mathcal{A}_t|}, \qquad \mathcal{G}_i \in \mathcal{A}_t, \tag{15}$$

samples, producing a balanced mixture of retained local distributions (Zhu et al., 2021). Each $\mathcal{G}_i$ samples $z \sim \mathcal{N}(0, I)$ to construct the server-side dataset D_t^{syn}. Starting from $\bar{\theta}^t$, the server performs masked reconstruction on D_t^{syn} using the same MIM objective, yielding refined parameters θ_{ref}^t. This reintroduces information from heterogeneous local distributions into the aggregated model through reliable sample-level surrogates.

Since D_t^{syn} remains an approximation of the original decentralized data, directly replacing $\bar{\theta}^t$ with θ_{ref}^t may amplify residual generation errors. FedMAGE therefore applies conservative parameter interpolation:

$$\theta^{t+1} = (1 - \rho)\bar{\theta}^t + \rho\, \theta_{\mathrm{ref}}^t, \tag{16}$$

where $\rho \in [0, 1]$ controls the refinement strength. The resulting θ^{t+1} preserves the consensus from federated aggregation while incorporating complementary information from reliable client generators, and is redistributed in the next round. This process repeats until federated pretraining ends. The complete FedMAGE procedure is summarized in Algorithm 1 in the appendix.

5 Experiments

Next, we empirically evaluate FedMAGE under non-IID federated settings and examine whether our theoretical findings can translate into practical gains in representation learning.

5.1 Experimental Setup

Benchmark Datasets. For federated self-supervised pretraining, we use Mini-ImageNet (Vinyals et al., 2016), which contains 60,000 images sampled from ImageNet (Deng et al., 2009). After pretraining, the learned representations are evaluated on CIFAR-10, CIFAR-100 (Krizhevsky, 2009), and ImageNet through downstream finetuning. CIFAR-10 and CIFAR-100 provide compact low-resolution benchmarks, while ImageNet offers a substantially larger and higher-resolution setting, allowing us to evaluate the learned representations across different data scales and visual complexities.

Federated Scenario. To model statistical heterogeneity among clients, we partition the pretraining data across $N = 100$ clients using Dirichlet sampling (Hsu et al., 2019), where a smaller concentration parameter α produces a more heterogeneous label distribution. At each communication round, 5 clients are randomly selected to participate in training. Each selected client performs 4 local epochs before returning its updated model to the server for aggregation.

Training Setup. FedMAGE is primarily implemented on a ViT-based MAE backbone (Dosovitskiy et al., 2020; He et al., 2022), while its design is also applicable to ResNet-based architectures. The variational generator uses a 128-dimensional latent space, with $\lambda = 0.5$ and $\beta = 0.01$ weighting the generative and KL objectives, respectively, while semantic and spatial consistency are combined using $\kappa = 0.7$ for quality estimation. Server-side synthetic refinement starts after $T_{\mathrm{w}} = 10$ rounds and is blended with the aggregated model using $\rho = 0.1$. Pretraining uses AdamW with a base learning rate of 1.5×10^{-4} and a batch size of 64 per GPU. All experiments are implemented in PyTorch on a server with 4 NVIDIA RTX 5090 GPUs, and complete hyperparameter and reproduction details are provided in Appendix A.2.

5.2 Empirical Evaluations

Comparison with Federated SSL Baselines. Our theoretical analysis provides representation-level insights into how generative learning can benefit federated SSL under heterogeneous data. However,

Table 1: **Top-1 finetuning accuracy (%) between representative F-SSL methods and FedMAGE under heterogeneous cross-device settings.** Results are reported as mean and standard deviation over three runs, with the best and second-best results highlighted in **bold** and underline, respectively.

Method	Architecture	Params	CIFAR-10	CIFAR-100	ImageNet
FedU	ResNet-18	38.47M	42.39 ± 0.21	30.98 ± 1.08	48.11 ± 2.32
FedEMA	ResNet-18	38.47M	33.97 ± 0.21	21.73 ± 0.16	41.28 ± 0.89
Orchestra	ResNet-18	11.84M	66.28 ± 0.75	46.85 ± 0.40	40.61 ± 0.82
FeatARC	ResNet-18	11.70M	57.07 ± 0.85	48.81 ± 1.04	57.91 ± 0.56
LDAWA	ResNet-18	15.39M	63.02 ± 0.15	33.62 ± 0.16	21.29 ± 0.69
FedU^2	ResNet-18	15.39M	42.66 ± 0.76	30.59 ± 1.24	35.39 ± 0.56
FedMAR	ResNet-18	22.50M	64.42 ± 0.65	39.47 ± 0.72	67.24 ± 0.38
FedMAGE (Ours)	ResNet-18	22.92M	$\underline{73.78} \pm 0.27$	$\underline{51.78} \pm 0.04$	$\underline{69.64} \pm 0.17$
FedMAGE (Ours)	Tiny-ViT	15.04M	$\mathbf{77.19} \pm 0.49$	$\mathbf{54.45} \pm 1.30$	$\mathbf{71.96} \pm 0.12$

Table 2: **Comparison under different data heterogeneity settings.** Table reports the average results (%) of three runs under label-skew non-IID (L-NonIID) and feature-skew non-IID (F-NonIID).

Method	Dataset	L-NonIID			F-NonIID
		$\alpha = 1$	$\alpha = 0.1$	$\alpha = 0.01$	
Federated MAE	CIFAR-10	60.28	59.41	59.38	61.37
	CIFAR-100	38.43	36.85	36.56	38.30
FedMAR	CIFAR-10	62.38	62.24	62.25	64.88
	CIFAR-100	37.23	37.02	36.94	39.98
FedMAGE	CIFAR-10	**77.17**	**77.19**	**76.61**	**77.04**
	CIFAR-100	**54.98**	**54.45**	**54.38**	**55.02**

given the gap between the analytical setting and practical scenarios, it remains necessary to examine whether these advantages persist empirically. Table 1 compares FedMAGE with state-of-the-art baselines addressing this challenge. Specifically, FedU (Zhuang et al., 2021a) and FedEMA (Zhuang et al., 2021b) adopt predictor- or momentum-based adaptation; Orchestra (Lubana et al., 2022), FeatARC (Wang et al., 2022), and FedMAR (Chen et al., 2026) use clustering or feature alignment for cross-client coordination; while LDAWA (Rehman et al., 2023) and FedU^2 (Liao et al., 2024) focus on adaptive aggregation to reduce representation drift. We use a highly non-IID setting with $\alpha = 0.1$, where all methods are pretrained for 200 communication rounds and then finetuned for 50 epochs on each downstream dataset.

As shown in Table 1, FedMAGE achieves the best performance on all three downstream benchmarks. With ResNet-18, FedMAGE outperforms all competing methods, reaching 73.78% on CIFAR-10 while increasing the parameter count only slightly from 22.50M in FedMAR to 22.92M. The gains remain consistent on CIFAR-100 and ImageNet, indicating that the benefit extends from compact low-resolution datasets to larger-scale visual recognition. To examine generalization beyond convolutional backbones, we also evaluate FedMAGE with Tiny-ViT, a lightweight Transformer (Vaswani et al., 2017; Dosovitskiy et al., 2020) of comparable model size to ResNet-18. FedMAGE with Tiny-ViT achieves even stronger performance across all three benchmarks while using only 15.04M parameters. These results suggest that the gains are not simply due to increased model capacity and remain consistent across backbone families and downstream data scales. More importantly, the improvements under severe heterogeneity empirically support our theoretical finding that generative enhancement can strengthen federated representation learning beyond conventional coordination.

Robustness Study under Various Forms of Data Heterogeneity. The previous comparison confirms the effectiveness of FedMAGE under a severe non-IID setting. We are then interested in whether this advantage generalizes across diverse heterogeneity settings. In particular, we evaluate label-skew settings with $\alpha \in \{1, 0.1, 0.01\}$ and a feature-skew setting. For label skew, smaller α indicates increasingly heterogeneous client label distributions. For feature skew, each client retains the full class distribution but applies a distinct data transformation, producing heterogeneous feature distributions without altering label coverage. To provide a focused comparison within federated MIM, we include

Table 3: **Ablation study on the key components of FedMAGE.** All variants are evaluated under the same federated setting. The accuracy results (%) are the average of three independent trials.

Variant	Local Gen.	Server Refine	Quality Guide	CIFAR-10	CIFAR-100
Vanilla Federated MAE	No	No	No	59.41	36.85
+ Local Generation	Yes	No	No	72.83	52.09
+ Server Refinement	Yes	Yes	No	75.32	54.36
FedMAGE	Yes	Yes	Yes	**77.19**	**54.45**

vanilla federated MAE and FedMAR (Chen et al., 2026), a recent method designed for federated masked representation learning. All other training configurations remain unchanged, and results are averaged over three runs. As shown in Table 2, FedMAGE achieves the best performance across both datasets and all non-IID settings. Its performance remains stable as label skew becomes more severe: on CIFAR-10, accuracy changes only from 77.17% at $\alpha = 1$ to 76.61% at $\alpha = 0.01$, with similarly small variation on CIFAR-100. FedMAGE also maintains a clear advantage under feature skew, showing that its benefit is not limited to Dirichlet-based label imbalance. Such robustness is particularly valuable in practical federated environments, where client distributions may vary substantially in both semantic composition and feature characteristics.

Ablation Study of FedMAGE. Next, to understand where the gains of FedMAGE originate, we progressively construct the framework from vanilla federated MAE by introducing local generative learning, server-side synthetic refinement, and quality-guided generator selection. The study uses the default non-IID setting, while all other setups remain unchanged, and the mean values of three trials are reported. Table 3 shows that local generative learning already brings substantial gains on both datasets, including an increase from 59.41% to 72.83% on CIFAR-10. Server-side refinement further improves performance, indicating that generated samples provide useful information beyond conventional parameter aggregation. With quality-guided generator selection, the complete FedMAGE achieves the best performance on both datasets. Overall, the gains arise from the complementary interaction between local and global generative mechanisms rather than any single component. They are also consistent with our theoretical analysis, where local generative learning strengthens semantic recovery at clients and reliable synthetic refinement enhances the aggregated representation.

Further Empirical Study. Beyond the above experiments, we provide additional evaluations in Appendix A.3 to examine more aspects of FedMAGE. We first compare alternative generator architectures and find that the lightweight variational design offers a favorable performance-complexity trade-off, while equal allocation of the fixed synthetic budget among reliable generators outperforms score-dependent strategies. We then examine representation spaces and parameter distances, where FedMAGE shows stronger consistency between heterogeneous local models and the aggregated global model. Finally, a generator memorization evaluation finds nearly identical similarities between generated samples and client training versus held-out data across training rounds, alleviating the concern that generative coordination may introduce substantial training-sample memorization risk.

6 Conclusion

This paper investigated how generative learning can improve federated self-supervised representation learning under heterogeneous client data. Our theoretical analysis shows that local generative learning strengthens globally relevant semantic recovery, while semantically consistent server-side refinement improves global representability. Guided by these important findings, we develop FedMAGE with local masked-generative learning and quality-guided synthetic refinement. Experiments show consistent gains over representative baselines, robustness to varying data heterogeneity, stronger local–global representation consistency, and clear benefits from its key design choices. Together, these results highlight generative coordination as a new avenue for federated SSL, enabling clients to contribute transferable information beyond parameter updates while preserving data locality.

7 AI USE STATEMENT

Generative AI tools were used to support several auxiliary aspects of this work, including checking and refining mathematical derivations and proof presentation, improving the clarity and readability of the manuscript, and adjusting the visual presentation of the framework figure. The core theoretical formulation, final mathematical arguments, experimental methodology, implementation, and empirical evaluation were developed and validated by the authors. All AI-assisted suggestions, including textual revisions, mathematical feedback, and visual modifications, were manually reviewed and revised before being incorporated into the paper. The authors take full responsibility for the final content, claims, and results presented in this work.

REFERENCES

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Varun Belagali, Srikar Yellapragada, Alexandros Graikos, Saarthak Kapse, Zilinghan Li, Tarak Nath Nandi, Ravi K Madduri, Prateek Prasanna, Joel Saltz, and Dimitris Samaras. Gen-sis: Generative self-augmentation improves self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2886–2896, 2026.

Chen Chen, Daochang Liu, and Chang Xu. Towards memorization-free diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8425–8434. IEEE, 2024.

Chen Chen, Daochang Liu, Mubarak Shah, and Chang Xu. Enhancing privacy-utility trade-offs to mitigate memorization in diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8182–8191. IEEE, 2025.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Xuanyu Chen, Nan Yang, Shuai Wang, and Dong Yuan. Understanding the robustness of distributed self-supervised learning frameworks against non-iid data. In *The Fourteenth International Conference on Learning Representations*, 2026.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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. In *International Conference on Learning Representations*, 2020.

Lele Fu, Sheng Huang, Yanyi Lai, Tianchi Liao, Chuanfu Zhang, and Chuan Chen. Beyond federated prototype learning: Learnable semantic anchors with hyperspherical contrast for domain-skewed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16648–16656, 2025. doi: 10.1609/aaai.v39i16.33829.

Georgi Ganev, Bristena Oprisanu, and Emiliano De Cristofaro. Robin hood and matthew effects: Differential privacy has disparate impact on synthetic data. In *International Conference on Machine Learning*, pp. 6944–6959. PMLR, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Ran He, Jie Cao, and Tieniu Tan. Generative artificial intelligence: a historical perspective. *National Science Review*, 12(5):nwaf050, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s): 1–37, 2022.

Shusen Jing, Anlan Yu, Shuai Zhang, and Songyang Zhang. Fedsc: Provable federated self-supervised learning with spectral contrastive objective over non-iid data. *arXiv preprint arXiv:2405.03949*, 2024.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2142–2152, 2023.

Xinting Liao, Weiming Liu, Chaochao Chen, Pengyang Zhou, Fengyuan Yu, Huabin Zhu, Binhui Yao, Tao Wang, Xiaolin Zheng, and Yanchao Tan. Rethinking the representation in federated unsupervised learning with non-iid data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22841–22850, 2024.

Hong Liu, Jeff Z HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. In *International Conference on Learning Representations*, 2022.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

Ying Liu, Jianhui Yin, Weiting Zhang, Changming An, Yu Xia, and Hongke Zhang. Integration of federated learning and ai-generated content: A survey of overview, opportunities, challenges, and solutions. *IEEE Communications Surveys & Tutorials*, 27(5):3308–3338, 2024.

Ekdeep Lubana, Chi Ian Tang, Fahim Kawsar, Robert Dick, and Akhil Mathur. Orchestra: Unsupervised federated learning via globally consistent clustering. In *International Conference on Machine Learning*, pp. 14461–14484. PMLR, 2022.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

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.

Kumar Kshitij Patel, Bingqing Jiang, AFM Kabir, Weitong Zhang, Difan Zou, and Lingxiao Wang. Personalized federated training of diffusion models with privacy guarantees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31790–31801, 2026.

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.

Fiona Victoria Stanley Jothiraj and Afra Mashhadi. Phoenix: A federated generative diffusion model. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 1568–1577, 2024.

Yan Sun, Li Shen, and Dacheng Tao. Towards understanding generalization and stability gaps between centralized and decentralized federated learning, 2024. URL `https://arxiv.org/abs/2310.03461`.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8432–8440, 2022. doi: 10.1609/aaai.v36i8.20819.

Zhenheng Tang, Shaohuai Shi, Bo Li, and Xiaowen Chu. Gossipfl: A decentralized federated learning framework with sparsified and adaptive communication. *IEEE Transactions on Parallel and Distributed Systems*, 34(3):909–922, 2022.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S Rellermeyer. A survey on distributed machine learning. *Acm computing surveys (csur)*, 53(2): 1–33, 2020.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Lirui Wang, Kaiqing Zhang, Yunzhu Li, Yonglong Tian, and Russ Tedrake. Does learning from decentralized non-iid unlabeled data benefit from self supervision? *arXiv preprint arXiv:2210.10947*, 2022.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9653–9663, 2022.

Baochen Xiong, Xiaoshan Yang, Yaguang Song, Yaowei Wang, and Changsheng Xu. Client-adaptive cross-model reconstruction network for modality-incomplete multimodal federated learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1241–1249, 2023a.

Zuobin Xiong, Wei Li, and Zhipeng Cai. Federated generative model on multi-source heterogeneous data in iot. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10537–10545, 2023b.

Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. A survey on federated learning. *Knowledge-Based Systems*, 216:106775, 2021.

Tianlong Zhang, Zhe Xue, Adnan Mahmood, Junping Du, Yuchen Dong, Shilong Ou, Lang Feng, Ming-Hsuan Yang, and Yuankai Qi. Generating synthetic data for unsupervised federated learning of cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22569–22577, 2025.

Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390, 2021.

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, 2021a.

Weiming Zhuang, Yonggang Wen, and Shuai Zhang. Divergence-aware federated self-supervised learning. In *International Conference on Learning Representations*, 2021b.

A Appendix

A.1 Training Procedure of FedMAGE

Algorithm 1 FedMAGE Algorithm

Input: initial global model θ^0, number of local epochs E, number of communication rounds T, learning rate η, loss weights λ and β, quality weight κ, warm-up rounds T_{w}, generator-quality threshold γ, generation budget B_{syn}, and interpolation weight ρ

Output: optimized global model θ^T

for $t = 0, \ldots, T-1$ **do**
  server samples participating client set $\mathcal{S}_t \subseteq [N]$ and broadcasts θ^t
  for client $i \in \mathcal{S}_t$ in parallel **do**
    initialize local model $\theta_i^t \leftarrow \theta^t$
    optimize θ_i^t for E local epochs using $\mathcal{L}_{\mathrm{local}}^{(i)}$ (Eq. 10)
    if $t \geq T_{\mathrm{w}}$ **then**
      generate reconstructions $\tilde{x}$ from local samples x using $\mathcal{G}_i$
      compute semantic and spatial consistency scores $\zeta_i^{\mathrm{sem}}(\tilde{x}, x)$ and $\zeta_i^{\mathrm{spa}}(\tilde{x}, x)$ (Eq. 11)
      compute combined quality score $\zeta_i(\tilde{x}, x)$ (Eq. 12)
      average sample-level scores to obtain $\bar{\zeta}_i$ and $\bar{\zeta}_i^{\mathrm{spa}}$
      send θ_i^t, $\bar{\zeta}_i$, and $\bar{\zeta}_i^{\mathrm{spa}}$ to server
    else
      send θ_i^t to server
    end if
  end for
  aggregate client models to obtain $\bar{\theta}^t$ (Eq. 13)
  if $t < T_{\mathrm{w}}$ **then**
    $\theta^{t+1} \leftarrow \bar{\theta}^t$
  else
    select reliable generator set $\mathcal{A}_t$ according to the quality threshold (Eq. 14)
    if $|\mathcal{A}_t| > 0$ **then**
      allocate B_i^t synthetic samples to each $\mathcal{G}_i \in \mathcal{A}_t$ (Eq. 15)
      construct D_t^{syn} by sampling $z \sim \mathcal{N}(0, I)$ from each retained generator
      initialize $\theta_{\mathrm{ref}}^t \leftarrow \bar{\theta}^t$
      refine θ_{ref}^t on D_t^{syn} using $\mathcal{L}_{\mathrm{MIM}}$ (Eq. 8)
      update the global model by conservative interpolation (Eq. 16)
    else
      $\theta^{t+1} \leftarrow \bar{\theta}^t$
    end if
  end if
end for

A.2 Implementation and Reproducibility Details

In this section, we provide additional implementation details and hyperparameter configurations to complement the experimental setup described in the main paper and assist reproducibility. Table 4 reports the hardware and software environment used to run the experiments, while Table 5 summarizes the federated setting, FedMAGE-specific configurations, and training parameters.

Table 4: **Hardware and software environment used in our experiments.**

Configuration	**Details**
Operating system	Ubuntu 22.04
Python	3.10
CUDA	11.8
Framework	PyTorch
GPU	4 $\times$ NVIDIA RTX 5090 (32 GB)
CPU	Intel Xeon Platinum 8470Q, 100 vCPUs
System memory	368 GB

Table 5: **Detailed experimental configuration of FedMAGE.** Default values are shown in **bold**, while the remaining values denote alternative settings used in robustness or sensitivity studies.

Configuration	Details
Federated Setup	
Pretraining dataset	Mini-ImageNet
Downstream datasets	CIFAR-10, CIFAR-100, ImageNet
Data partition	Dirichlet / feature skew
Dirichlet coefficient α	1, **0.1**, 0.01
Number of clients N	**100**
Clients sampled per round	**5**
Local training epochs	**4**
Communication rounds	100, **200**
Input resolution	224×224
FedMAGE Configuration	
SSL objective	Masked image modeling (MAE)
Backbone	**Tiny-ViT**, ResNet-18
Patch size	14×14
Transformer encoder depth / dimension	**9 / 320**
Transformer decoder depth / dimension	**1 / 128**
Attention heads	**16**
Mask ratio	**0.75**
Generator	Variational image generator
Latent dimension	**128**
Generative loss weight λ	0.25, **0.5**
KL weight β	**0.01**
Semantic weight κ	**0.7**
Combined confidence threshold	**0.75**, 1.0
Spatial confidence threshold	**0.60**, 1.0
Prior evaluation samples/client	**128**
Reference batches/client	**4**
Generator updates encoder	False, **True**
Server generator quality threshold γ	**0.50**
Synthetic budget B_{syn}	**512** samples/round
Server refinement epochs	**1**
Refinement start round T_{w}	**10**
Global interpolation coefficient ρ	**0.10**
Training Details	
Optimizer	AdamW
Base pretraining learning rate	1.5×10^{-4}
Weight decay	**0.05**
Pretraining batch size	**64**/GPU, 128
Pretraining gradient accumulation	**1**
Finetuning batch size	**128**/GPU, 512
Finetuning gradient accumulation	**2**
Finetuning epochs	**50**
Finetuning peak learning rate	4×10^{-3}
Finetuning LR schedule	5-epoch warm-up + cosine decay
Layer-wise LR decay	**0.75**
Random seeds	0, 1, 2
Evaluation metric	Top-1 accuracy (%)

A.3 Further Experiments

Generator Choice Analysis. In addition to the main evaluations, we continue to explore if the effectiveness of FedMAGE depends on the specific choice of generative architecture. Since VAEs (Kingma & Welling, 2014), GANs (Goodfellow et al., 2020), and diffusion models (Ho et al., 2020) represent three widely adopted paradigms for generative modeling, we additionally instantiate FedMAGE with GAN- and diffusion-based generators as representative alternatives to the default variational design. For a controlled comparison, only the generator architecture is replaced, while the remaining framework, backbone, and federated training configuration are kept unchanged, and all variants are evaluated under the default experimental setting. The results in Table 6 show that all three generator architectures achieve comparable downstream performance, suggesting that the effectiveness of FedMAGE is not tied to a particular generative model family. Among them, the VAE-based design achieves the best accuracy on both CIFAR-10 and CIFAR-100 while requiring the fewest parameters and substantially lower computational cost, with only 7.66 GFLOPs compared with 18.34 and 22.98 GFLOPs for the GAN- and diffusion-based variants, respectively. This indicates that adopting a more complex generator is unnecessary for obtaining effective synthetic refinement in our setting. More broadly, the consistent performance across distinct generative paradigms suggests that the gains of FedMAGE mainly stem from its generative coordination mechanism rather than a particular generator architecture, while the lightweight variational design offers the most favorable balance between downstream performance, model size, and computational efficiency among the compared variants.

Table 6: **Effect of different generator architectures in FedMAGE.** We report the mean Top-1 finetuning accuracy (%) over three runs under the same federated setting.

Generator	Latent/Input Type	Total Params	GFLOPs	CIFAR-10	CIFAR-100
GAN	Gaussian latent	15.73M	18.34	76.85	54.36
Diffusion	Gaussian noise	15.32M	22.98	77.06	54.30
VAE (FedMAGE)	Gaussian latent	15.04M	7.66	**77.19**	**54.45**

Synthetic Budget Allocation Analysis.

Following the clarification of our generator choice, we next examine the server-side refinement design, which constitutes another important component of FedMAGE. As the synthetic training set used for refinement is jointly constructed from multiple reliable client generators, the way in which the fixed budget B_{syn} is distributed across these generators directly determines the composition of the server data. To characterize this design choice, we evaluate three allocation strategies under the same total budget and default federated setting. Quality-Proportional allocation uses $B_i = B_{\text{syn}}\zeta_i / \sum_{j\in\mathcal{A}_t}\zeta_j$, so that generators with higher quality scores contribute more synthetic samples. Focal-Adjusted allocation starts from the equal-share budget $B_{\text{syn}}/|\mathcal{A}_t|$ and perturbs it using a centered focal-style term based on $(1-\zeta_i)^2$, allowing lower-confidence but still admissible generators to receive a moderately larger share while preserving the same total budget. FedMAGE instead assigns the budget uniformly as $B_i = B_{\text{syn}}/|\mathcal{A}_t|$ after quality-based filtering. Table 7 shows that both score-dependent alternatives lead to lower downstream performance than equal allocation. On CIFAR-10, the accuracy decreases from 77.19% with equal allocation to 75.84% with Focal-Adjusted allocation and 74.64% with Quality-Proportional allocation. A similar trend appears on CIFAR-100, where the accuracy drops from 54.45% to 53.29% and 52.08%, respectively. These consistent drops imply that, once unreliable generators have been removed, quality scores are better suited for determining generator reliability than for continuously controlling synthetic-data volume. Biasing the budget toward particular clients can overemphasize portions of the heterogeneous local distributions, whereas equal allocation preserves a more balanced cross-client contribution and yields a synthetic mixture that more closely reflects the distributed global data composition. Accordingly, FedMAGE uses quality scores to determine which generators participate in refinement, while allocating the synthetic budget uniformly among the selected generators.

Table 7: **Comparison of synthetic budget allocation strategies.** All strategies use the same total generation budget B_{syn}. The average finetuning accuracies (%) over three trials are reported.

Allocation Strategy	Allocation Rule	CIFAR-10	CIFAR-100
Quality-Proportional	$B_i = B_{\text{syn}} \frac{\zeta_i}{\sum_{j \in \mathcal{A}_t} \zeta_j}$	74.64	52.08
Focal-Adjusted	$B_i = \frac{B_{\text{syn}}}{\lvert\mathcal{A}_t\rvert} \left[1 + (1-\zeta_i)^2 - \frac{1}{\lvert\mathcal{A}_t\rvert} \sum_{j \in \mathcal{A}_t} (1-\zeta_j)^2 \right]$	75.84	53.29
Equal (FedMAGE)	$B_i = \frac{B_{\text{syn}}}{\lvert\mathcal{A}_t\rvert}$	**77.19**	**54.45**

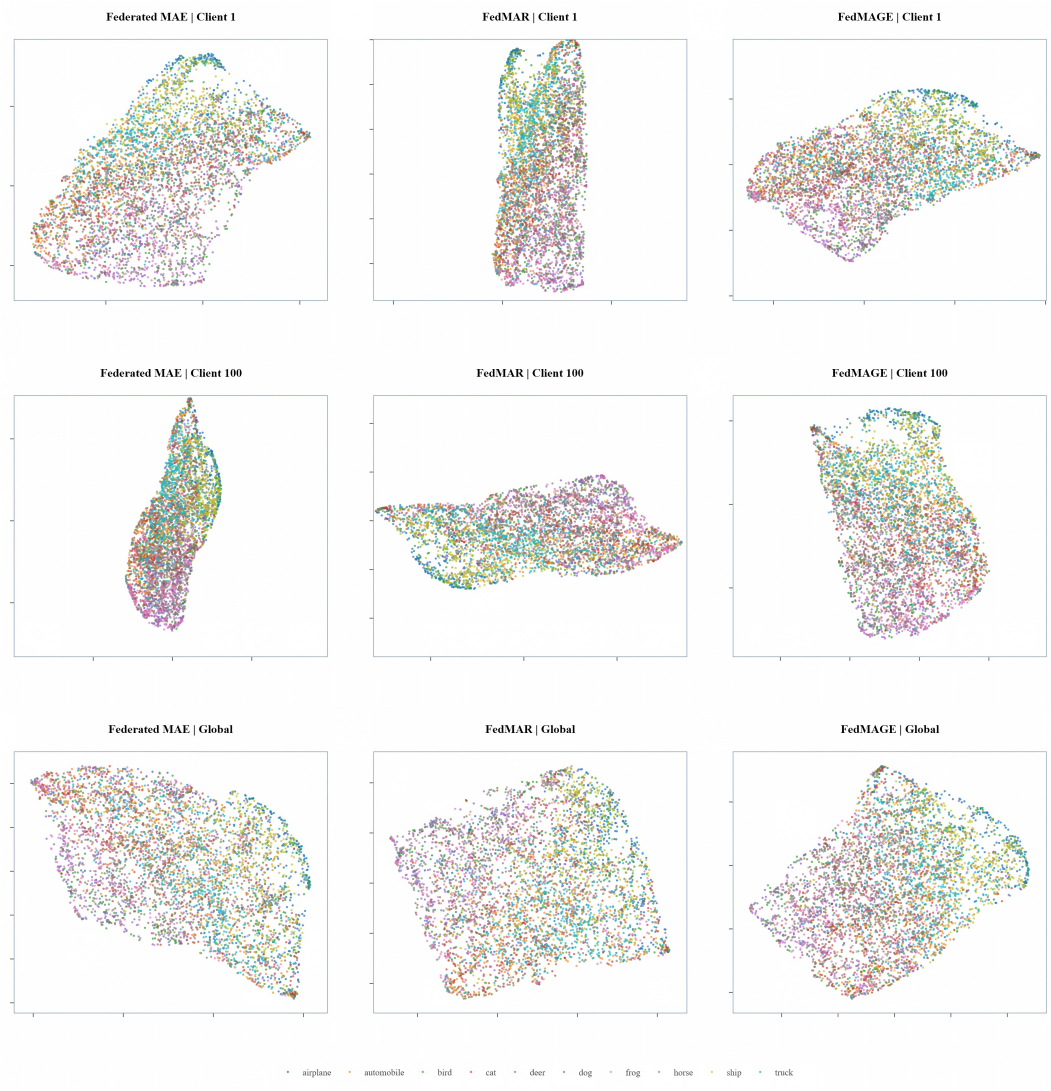

Figure 3: **Representation spaces learned by different federated masked representation learning methods.** The three columns correspond to Federated MAE, FedMAR, and FedMAGE, respectively. From top to bottom, the rows visualize the feature spaces of the local models independently trained on the data of Client 1 and Client 100, followed by that of the aggregated global model. Colors denote different semantic classes.

Table 8: **Cross-model parameter consistency under heterogeneous federated training.** We report the summed layer-wise ℓ_2 distance between representative local models and the aggregated global model. Lower values indicate stronger consistency across heterogeneous clients.

Method	Local_1-Local_{100}	Local_1-Global	Local_{100}-Global	Avg. Local–Global
Federated MAE	88.44	72.19	80.10	76.15
FedMAR	79.00	79.44	82.19	80.82
FedMAGE	**77.34**	**68.83**	**66.76**	**67.80**

Table 9: **Generated-sample similarity to client training and held-out data across communication rounds.** Train NN Sim. and Holdout NN Sim. denote the average maximum cosine similarities to the corresponding reference subsets, and $\Delta_{\text{mem}} = S_{\text{train}} - S_{\text{hold}}$.

Round	Train NN Sim.	Holdout NN Sim.	Δ_{mem}
50	0.9593	0.9588	0.0005
100	0.8981	0.8997	-0.0016
200	0.8268	0.8276	-0.0007

Representation Space Consistency Study. The non-IID evaluation reported in the main manuscript shows that FedMAGE maintains strong downstream performance across different forms of client heterogeneity. However, downstream utility alone provides only a task-level view and does not reveal how the learned representations themselves evolve under heterogeneous local data. A more direct inspection of the feature space is therefore needed to examine whether the observed empirical gains are accompanied by the representation-level behavior suggested by our analysis. We conduct this representation-space study under the same default federated setting as the main experiments, with 100 clients and Dirichlet sampling at $\alpha = 0.1$. To ensure a fair comparison, we restrict the study to federated MIM approaches with closely related pretraining objectives. Specifically, we include Federated MAE as the standard federated MIM baseline, FedMAR as a recent heterogeneity-aware federated MIM method, and FedMAGE as our generative-enhanced approach. For each method, we retain two representative local models trained independently using only the data of Client 1 and Client 100, together with the aggregated global model. We extract encoder features from the same CIFAR-10 evaluation samples, reduce them to 20 dimensions using principal component analysis (PCA), and then embed them into two dimensions using UMAP (McInnes et al., 2018). Since low-dimensional projections offer an intuitive but inherently qualitative view of representation structure, we additionally quantify model consistency by measuring the parameter distances between the two local models and between each local model and the corresponding global model.

Figure 3 presents the resulting representation spaces. All three methods learn meaningful semantic structures, while FedMAR and FedMAGE exhibit more coherent organization across heterogeneous local models. Among them, FedMAGE shows the strongest consistency between the two local representations and the aggregated global representation. The quantitative results in Table 8 provide complementary evidence from the parameter space. Compared with Federated MAE, FedMAR reduces the Local_1–Local_{100} distance from 88.44 to 79.00, while FedMAGE decreases it to 77.34 and also achieves the smallest local-global distances for Client 1 and Client 100, at 68.83 and 66.76, respectively. Its average local-global distance reaches 67.80, the lowest among the compared methods. Together, these results show that the gains of FedMAGE are not limited to downstream accuracy. Its generative coordination also promotes more consistent representations across heterogeneous clients and brings local models into closer agreement with the aggregated global model, in line with the representation-level behavior characterized by our theoretical analysis.

Generated-Data Memorization Risk Evaluation. FedMAGE relies on client-side generators to synthesize additional data for server-side refinement, allowing client knowledge to contribute to global learning without directly sharing raw local observations. Nevertheless, generative models can potentially retain excessive information about their training samples, raising the concern that generated data may become overly similar to specific local observations (Hu et al., 2022; Chen et al., 2025; 2024). This aspect has not been explicitly examined in our preceding evaluations, motivating us to investigate the memorization behavior of the client generators. We conduct the evaluation at communication rounds 50, 100, and 200. After constructing the non-IID client partitions, the local

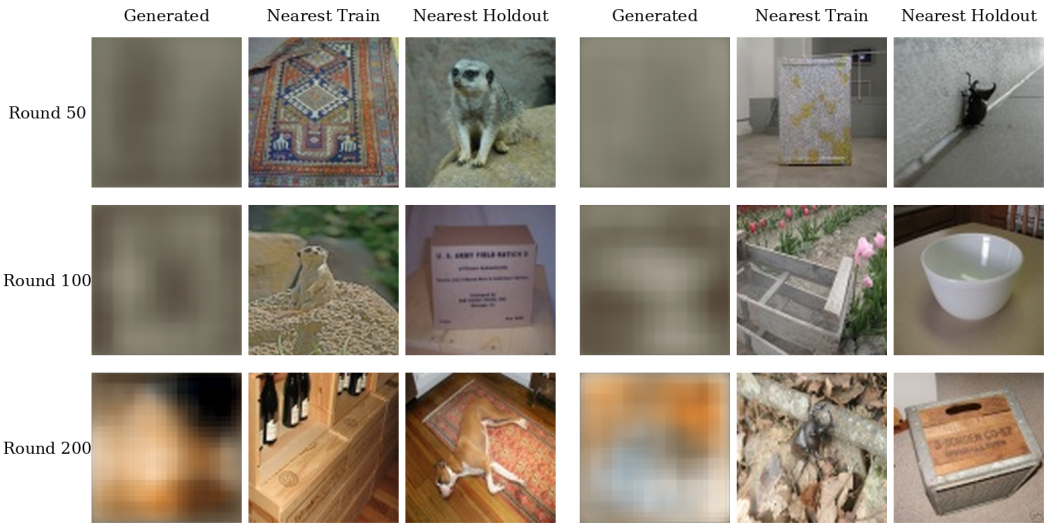

Figure 4: **Qualitative nearest-neighbor comparison of generated samples.** Rows correspond to Rounds 50, 100, and 200, respectively. Each row contains two generated examples, and each example is paired with its nearest training and held-out samples in the pretrained feature space.

data of each evaluated client are further divided equally into a training subset and a held-out subset, with only the former used to optimize the corresponding generator. Synthetic samples are then drawn from the random prior, following the same generation mode used by FedMAGE for server-side refinement. For each generated sample $\tilde{x}$, we extract its representation using a fixed pretrained feature encoder and separately identify its nearest neighbor from the training and held-out subsets using cosine similarity. Their average similarities are denoted by S_{train} and S_{hold}, respectively, and we measure their difference as $\Delta_{\text{mem}} = S_{\text{train}} - S_{\text{hold}}$. A larger positive gap would indicate a stronger preference toward previously observed training instances, whereas a value close to zero reflects comparable proximity to seen and unseen samples.

Table 9 reports the quantitative results across the three communication rounds. The nearest-neighbor similarities to training and held-out samples remain closely matched throughout training. At Round 50, S_{train} and S_{hold} are 0.9593 and 0.9588, respectively, yielding a gap of only 0.0005. The corresponding gaps at Rounds 100 and 200 remain similarly negligible at -0.0016 and -0.0007. These consistently small differences indicate that the generated samples do not exhibit a systematic preference toward the local data used for generator training. Interestingly, the absolute nearest-neighbor similarities to both subsets also decrease consistently as communication proceeds, from approximately 0.96 at Round 50 to 0.83 at Round 200. Since this reduction occurs simultaneously for both training and held-out references while their gap remains close to zero, the later generators appear less concentrated around individual local reference samples rather than shifting their preference from one subset to the other. This behavior is consistent with the intended role of FedMAGE, where generative learning is used to capture transferable information from local data for global refinement instead of reproducing particular training instances. For a more intuitive view of this learning behavior, Figure 4 presents generated examples together with their nearest counterparts from the training and held-out subsets across different communication rounds. The visual comparisons show that the nearest training and held-out samples exhibit comparable color distributions and coarse structural patterns relative to the generated outputs, without revealing a clear training-specific match. This observation is consistent with the numerical results and provides a qualitative view of the limited instance-level dependence of the generated data. Overall, these results suggest that FedMAGE can exploit client-side generative knowledge for global refinement while showing limited training-instance dependence under this evaluation, alleviating concerns that its effectiveness primarily originates from memorizing and reproducing local observations.

A.4 DISCUSSIONS

Communication Overhead. Since FedMAGE introduces generative coordination beyond conventional federated MIM, it is important to examine whether this design substantially increases communication cost. FedMAGE preserves the standard client-server model aggregation procedure, while introducing a lightweight generation branch and compact generator-quality statistics for server-side synthetic refinement.

The main communication increase comes from the additional parameters introduced by the generation branch. To quantify this cost, we additionally measure the parameter size of the corresponding federated MIM models using the same backbones but without the generative component. With ResNet-18 as the model backbone, the original federated MIM model contains approximately 22.50M parameters, whereas FedMAGE contains 22.92M parameters. The generation-related component therefore introduces only about 0.42M additional parameters, corresponding to

$$\frac{22.92 - 22.50}{22.50} \approx 1.9\% \tag{17}$$

of the original model size. Under FP32 transmission, this amounts to approximately 1.68MB of additional communication per transmission, compared with roughly 90MB for the original model. Similarly, with Tiny-ViT, the parameter count increases from 14.60M to 15.04M, corresponding to

$$\frac{15.04 - 14.60}{14.60} \approx 3.0\%, \tag{18}$$

or approximately 1.76MB of additional FP32 communication relative to the original 58.4MB model. Thus, the additional parameter communication introduced by FedMAGE remains around 3% or less for both evaluated backbones.

Besides model parameters, each participating client transmits two scalar generator-quality statistics, $\bar{\zeta}_i$ and $\bar{\zeta}_i^{\mathrm{spa}}$, after the warm-up stage. These statistics are locally averaged before transmission, so only two floating-point values are uploaded per sampled client. Under FP32 representation, this corresponds to only 8 bytes per client. With five clients sampled in each communication round, the total additional uplink communication introduced by quality evaluation is merely 40 bytes per round, which is negligible compared with model transmission.

In general, FedMAGE introduces only a modest increase in communication overhead: the generation branch increases model transmission by approximately 1.9%–3.0%, while the additional quality statistics contribute effectively negligible overhead. Moreover, the subsequent synthetic generation and refinement are performed only at the server, introducing no further client-server communication. Therefore, the additional coordination introduced by FedMAGE can be realized without substantially changing the communication pattern of conventional federated MIM.

Privacy Implications of Generative Coordination. Privacy is another important consideration for FedMAGE, as its generative coordination introduces additional information into the conventional FL process. Despite this extension, FedMAGE preserves the fundamental data-locality principle of FL: raw client images and individual training samples remain stored locally and are never transmitted to the server. The server receives only the updated model parameters together with compact generator-quality statistics, while the synthetic samples used for global refinement are generated from the learned local generators rather than directly collected from client datasets.

The additional quality information introduced by FedMAGE is also highly compressed. Specifically, each participating client uploads only the locally averaged statistics $\bar{\zeta}_i$ and $\bar{\zeta}_i^{\mathrm{spa}}$, which summarize the semantic and spatial reliability of its generator. These quantities are aggregated over generated candidates and therefore do not correspond to individual images, patch features, or sample-level embeddings. As a result, generator selection can be performed without requiring the server to access fine-grained representations of local training data.

A remaining concern is that generative models may memorize characteristics of their training samples, potentially causing generated outputs to reveal instance-level information (Hu et al., 2022). We empirically examine this issue through the generated-data memorization evaluation reported in Appendix A.3. Across multiple communication rounds, generated samples exhibit nearly identical nearest-neighbor similarities to seen training data and unseen held-out data, with the train-holdout similarity gap remaining close to zero. The absolute similarities also decrease as training proceeds,

suggesting that later generators become less concentrated around individual local reference samples. These observations provide empirical evidence that the synthetic information used by FedMAGE is not primarily obtained by reproducing specific training instances, which helps alleviate concerns about training-sample memorization in the generative coordination process.

Furthermore, for applications requiring stronger privacy protection, FedMAGE remains compatible with established privacy-enhancing mechanisms. Differential privacy can be incorporated during local optimization, for example by clipping and perturbing gradients of the encoder or generation branch to limit the influence of individual samples (Ganev et al., 2022; Patel et al., 2026). Complementary safeguards developed specifically for generative models, such as detecting or suppressing generated samples that are excessively similar to training instances (Chen et al., 2024; 2025), can also be integrated into local training or server-side sample screening. These mechanisms can strengthen privacy protection without altering the overall coordination structure of FedMAGE.

Practical Scope and Limitations. Beyond the communication and privacy considerations discussed above, it is also useful to clarify the practical scope of the current study. This work takes a step toward integrating generative learning with federated self-supervised representation learning, but several design choices were deliberately made to balance analytical tractability, training stability, and federated deployment cost. These choices define the current scope of FedMAGE and also point to several natural directions for future extension.

First, FedMAGE currently adopts a lightweight variational generator as its generative component. This choice is motivated by the constraints of federated training, where communication efficiency and local optimization stability are both important. Compared with larger generative models such as GANs (Goodfellow et al., 2020) or diffusion models (Ho et al., 2020), the VAE-based design is easier to optimize jointly with the local MIM objective and introduces only a small parameter increase, as discussed above. Our ablation results also provide empirical support for this design choice under the considered training budget. Nevertheless, stronger generative models may offer higher sample fidelity when sufficient computational resources are available. A particularly promising direction is to leverage pretrained generators and adapt only lightweight components during federated training, which may retain the benefit of stronger generative priors without incurring the full cost of training large generators from scratch.

Apart from the choice of generator, the current theoretical analysis and algorithmic design are developed specifically for federated masked image modeling. This focus is motivated by the favorable robustness of MIM under heterogeneous data observed in prior work (Chen et al., 2026), together with the natural overlap between masked reconstruction and generative modeling. These properties make MIM a suitable setting for studying how generative learning can improve both local representation recovery and global coordination. However, the same principle has not yet been formally established for other self-supervised paradigms, such as contrastive learning (Liu et al., 2021). Extending the theoretical analysis and generative coordination mechanism to contrastive or multimodal self-supervised learning therefore represents an important direction for future work.

Finally, our theoretical analysis adopts a simplified latent-class model to isolate and characterize client-specific semantic heterogeneity in a tractable manner. While this abstraction enables rigorous analysis and has been widely adopted in previous theoretical studies (Wang et al., 2022; Chen et al., 2026), real federated environments may exhibit more complex forms of statistical and system heterogeneity, including class imbalance, domain shift, unequal data volumes, and varying client availability. We complement the theoretical setting with experiments covering both Dirichlet-based label skew under different levels of non-IID data and feature-skew settings, providing additional empirical evidence beyond the analytical construction. Future work could further examine FedMAGE under compound statistical and system heterogeneity to better characterize its behavior in more diverse deployment scenarios.

A.5 NOTATION REFERENCE

To facilitate the understanding of mathematical notation, we organize the main symbols used in this paper into three groups. Table 10 summarizes the notation shared across the problem setup and learning objectives. Table 11 collects symbols introduced specifically for the theoretical analysis, while Table 12 shows the notation used in introducing FedMAGE.

Symbol	Meaning
$N,\ i \in [N]$	Number of clients and client index
D_i	Local dataset maintained by client i
$\mathcal{D}_i$	Local data distribution of client i
$\mathcal{D}$	Global data distribution, $\mathcal{D} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{D}_i$
d	Ambient dimension of the input space
c	Dimension of the learned representation
x	Input sample in $\mathbb{R}^d$
e_k	k-th standard basis vector of $\mathbb{R}^d$
$q^{(\cdot,\cdot)}$	Bernoulli variables controlling cross-client semantic interference
τ	Magnitude of cross-client semantic interference
μ	Scale of sample-level Gaussian perturbation
ξ	Gaussian noise vector with $\xi \sim \mathcal{N}(0, I_d)$
m	Binary mask with $m_j = 1$ for visible and $m_j = 0$ for masked coordinates
p_m	Per-coordinate masking probability, $\Pr(m_j = 0) = p_m$
x^{vis}	Visible input component, $x \odot m$
x^{mask}	Masked input component, $x \odot (1 - m)$
λ	Weight of the local generative objective
β	Weight of the KL regularization term

Table 10: **Basic notation for the federated system, data model, and masking process.**

Symbol	Meaning
W	Linear encoder matrix in $\mathbb{R}^{c\times d}$
V	Linear decoder matrix used in the theoretical analysis
$q_W(z \mid x)$	Variational posterior induced by encoder W
σ^2	Isotropic latent variance of the variational posterior
$\mathcal{L}_{\mathrm{MIM}}^{(i)}$	Local masked image modeling objective of client i
$\mathcal{L}_{\mathrm{VAE}}^{(i)}$	Local variational generative objective of client i
$\mathcal{L}_{\mathrm{loc}}^{(i)}$	Joint local objective combining masked reconstruction and generation
$\mathcal{L}_{\mathrm{Fed}}$	Federated objective obtained by aggregating local objectives
$\mathcal{L}_{\mathrm{srv}}$	Server-side masked reconstruction objective on synthetic data
$\mathcal{L}_{\mathrm{Global}}$	One-round surrogate objective combining $\mathcal{L}_{\mathrm{Fed}}$ and $\mathcal{L}_{\mathrm{srv}}$
A_i	Visible-input second-order statistic induced by random masking
B_i	Cross-statistic between visible and masked input components
Σ_i	Second-order statistic of the complete local data distribution
H_i	Combined local input statistic, $H_i = A_i + \Sigma_i$
G_i	Combined local cross-statistic, $G_i = B_i + \Sigma_i$
$\bar{W}, \bar{V}$	Aggregated encoder and decoder used in the global analysis
$\widetilde{\mathcal{D}}_i$	Synthetic distribution generated from client i
$\widetilde{\mathcal{D}}$	Global synthetic mixture formed from local generators
$\bar{\Sigma}$	Global second-order statistic of the original data distribution
$\widetilde{\Sigma}$	Second-order statistic of the synthetic distribution
$\mathcal{S}$	Global semantic subspace, $\mathrm{span}\{e_1, \ldots, e_N\}$
$P_{\mathcal{S}}$	Orthogonal projector onto $\mathcal{S}$
S_d	Weakest population semantic strength in the global semantic subspace
ω_d	Relative approximation factor in the synthetic semantic consistency condition
$r(W)$	Representability vector of encoder W
$r_k(W)$	Representability score of semantic direction e_k
$\delta_d^{\mathrm{MIM}}, \delta_d^{\mathrm{Gen}}, \delta_d^{\mathrm{Global}}, \delta_d$	Population-to-empirical perturbation terms

Table 11: **Notation used in the theoretical characterization of local generative learning and server-side synthetic refinement.**

Symbol	Meaning
t	Communication round index
$\mathcal{S}_t$	Set of participating clients at round t
θ_i^t	Model parameters returned by client i at round t
$\bar{\theta}^t$	Global parameters obtained after client model aggregation
w_i^t	Normalized aggregation weight of client i
$E_i, R_i, \mathcal{G}_i$	Client encoder, masked reconstruction decoder, and image generator
Ω	Set of masked patch indices
h_i^{cls}	`[CLS]` representation used by the generative branch
$Q_i(z \mid x)$	Client-specific variational posterior
ν_i, ℓ_i, z_i	Mean, log-variance, and reparameterized latent variable of $Q_i(z \mid x)$
$\tilde{x}_i$	Image synthesized by client generator $\mathcal{G}_i$
$\mathcal{L}_{\mathrm{gen}}^{(i)}$	Local variational generation objective
$\mathcal{L}_{\mathrm{local}}^{(i)}$	Joint local masked-generative training objective
$\phi_i^{\mathrm{sem}}(\cdot), \phi^{\mathrm{spa}}(\cdot)$	Semantic encoder feature and spatial pooled feature used for quality evaluation
$\zeta_i^{\mathrm{sem}}, \zeta_i^{\mathrm{spa}}, \zeta_i$	Semantic, spatial, and combined generation-quality scores
κ	Weight balancing semantic and spatial consistency
$\bar{\zeta}_i, \bar{\zeta}_i^{\mathrm{spa}}$	Candidate-averaged combined and spatial quality scores
T_{w}	Number of warm-up communication rounds
γ	Server-side generator-quality threshold
$\mathcal{A}_t$	Set of reliable generators selected at round t
B_{syn}	Total synthetic generation budget at each refinement round
B_i^t	Generation budget assigned to retained generator $\mathcal{G}_i$
D_t^{syn}	Synthetic dataset constructed at the server in round t
θ_{ref}^t	Parameters after server-side synthetic refinement
ρ	Conservative interpolation weight between aggregated and refined models
θ^{t+1}	Global model redistributed in the next communication round

Table 12: **FedMAGE-specific notation for local masked-generative learning, generator quality evaluation, and server-side refinement.**

A.6 Theoretical Analysis

A.6.1 Complete Problem Setup

Heterogeneous Data Modeling. We study a federated system with N clients, where client $i \in [N]$ holds a private unlabeled dataset D_i drawn from a client-specific distribution $\mathcal{D}_i$. All clients contain the same number of samples, i.e., $|D_1| = \cdots = |D_N|$, and their distributions jointly form the global distribution $\mathcal{D} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{D}_i$. Following prior theoretical studies of distributed SSL under non-IID data (Wang et al., 2022; Liu et al., 2022; Chen et al., 2026), we adopt a simplified latent-class construction to capture client-specific semantic imbalance while keeping the analysis tractable. Specifically, the global data space contains $2N$ latent classes, whereas each local distribution concentrates most of its mass on only two client-specific classes. Classes $2i-1$ and $2i$ are dominant on client i, contain the same number of samples, and both grow polynomially with the ambient dimension d. The remaining samples are sparsely distributed over classes in $[2N] \setminus \{2i-1, 2i\}$, with their aggregate amount growing only sublinearly as $O(d^\alpha)$ for some $\alpha \in (0, 1)$. Such an asymmetric allocation makes the dominant semantics statistically prevailing on each client, while the collection of local datasets jointly retains coverage of the global semantic space. To make this client-specific semantic concentration analytically identifiable in the input space, let $e_1, \ldots, e_d$ be the standard basis of $\mathbb{R}^d$, with d sufficiently large. The two dominant classes on client i are centered around opposite client-specific directions,

$$x^{(2i-1)} = e_i - \sum_{k \neq i} q^{(2i-1,k)} \tau e_k + \mu \xi^{(2i-1)}, \tag{19}$$

and

$$x^{(2i)} = -e_i - \sum_{k \neq i} q^{(2i,k)} \tau e_k + \mu \xi^{(2i)}, \tag{20}$$

where $q^{(\cdot,\cdot)}$ determines whether a cross-client semantic component is present and are independently sampled from $\mathrm{Bernoulli}(1/2)$, $\tau > 0$ controls its magnitude, and $\xi^{(\cdot)} \sim \mathcal{N}(0, I_d)$ introduces sample-level variation. Together with the polynomial-scale dominance of classes $2i-1$ and $2i$ and the sublinear presence of the remaining classes, this construction yields a controlled heterogeneous setting in which local training is primarily governed by client-specific semantic directions while the federation still spans the global class space.

Client-Side Learning Objective. Under the heterogeneous setting described above, each client optimizes the same model architecture using only its local distribution $\mathcal{D}_i$. We model the network with a shared linear encoder $W \in \mathbb{R}^{c \times d}$ and a generative decoder $V \in \mathbb{R}^{d \times c}$. The decoder is trained for both masked reconstruction and complete-data generation, while the same encoder W is shared by the two objectives so that both learning signals directly act on a common representation space. Given an input x, a random binary mask $m \in \{0,1\}^d$ partitions it into the visible and masked components,

$$x^{\mathrm{vis}} = x \odot m, \qquad x^{\mathrm{mask}} = x \odot (1 - m), \tag{21}$$

where each entry of m is independently set to 0 with probability p_m and to 1 with probability $1 - p_m$. The encoder maps x^{vis} into the representation space and the decoder reconstructs the missing content, giving the local masked reconstruction objective

$$\mathcal{L}^{(i)}_{\mathrm{MIM}}(W, V) = \mathbb{E}_{x \sim \mathcal{D}_i} \mathbb{E}_m \left[\left\| VWx^{\mathrm{vis}} - x^{\mathrm{mask}} \right\|_2^2 \right]. \tag{22}$$

To enable the same encoder–decoder pair to additionally model complete local samples, we instantiate the generative component as a linear-Gaussian variational autoencoder (VAE). For a complete input x, the shared encoder parameterizes the approximate posterior

$$q_W(z \mid x) = \mathcal{N}\left(Wx, \sigma^2 I_c\right), \tag{23}$$

where $\sigma^2 > 0$ is a fixed isotropic latent variance, and the prior is $p(z) = \mathcal{N}(0, I_c)$. A latent variable $z \sim q_W(z \mid x)$ is decoded by V to reconstruct the complete input. The corresponding generative objective on client i is

$$\mathcal{L}^{(i)}_{\mathrm{VAE}}(W, V) = \mathbb{E}_{x \sim \mathcal{D}_i} \mathbb{E}_{z \sim q_W(z|x)} \left[\|Vz - x\|_2^2 \right] + \beta \mathbb{E}_{x \sim \mathcal{D}_i} \left[D_{\mathrm{KL}} \left(q_W(z \mid x) \,\|\, p(z) \right) \right]. \tag{24}$$

Combining the two learning objectives, the client-side optimization is defined as

$$\mathcal{L}_{\mathrm{loc}}^{(i)}(W,V) = \mathcal{L}_{\mathrm{MIM}}^{(i)}(W,V) + \lambda \mathcal{L}_{\mathrm{VAE}}^{(i)}(W,V), \tag{25}$$

where $\lambda > 0$ controls the contribution of generative learning and $\beta > 0$ weights the latent regularization. To simplify the subsequent theoretical analysis, we adopt the normalized setting $\lambda = \beta = 1$ unless otherwise specified. The resulting objective characterizes the local model optimized on client i before federated aggregation.

Federated Global Optimization. After local optimization, the server aggregates the client models to obtain the global parameters. Since all clients have equal-sized local datasets, we adopt uniform aggregation,

$$\bar{W} = \frac{1}{N}\sum_{i=1}^{N} W_i, \qquad \bar{V} = \frac{1}{N}\sum_{i=1}^{N} V_i. \tag{26}$$

In standard federated training, the global model is determined solely by optimization over the heterogeneous client distributions. The corresponding federated objective can be characterized as

$$\mathcal{L}_{\mathrm{Fed}}(W,V) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\mathrm{loc}}^{(i)}(W,V). \tag{27}$$

The generative decoders learned during local training additionally allow the server to construct synthetic data carrying complementary client-specific information. For each client i, we sample

$$z_i \sim \mathcal{N}(0, I_c) \tag{28}$$

from the VAE prior and generate a complete synthetic sample using its locally optimized decoder,

$$\tilde{x}_i = V_i z_i. \tag{29}$$

Let $\widetilde{\mathcal{D}}_i$ denote the distribution induced by $V_i z_i$. By drawing the same number of synthetic samples from each local decoder, the resulting server-side distribution is represented by the uniform mixture

$$\widetilde{\mathcal{D}} = \frac{1}{N}\sum_{i=1}^{N} \widetilde{\mathcal{D}}_i. \tag{30}$$

Once generated, $\widetilde{\mathcal{D}}$ is treated as fixed during server-side optimization. The aggregated model is trained on these synthetic samples using the same masking mechanism defined above, leading to

$$\mathcal{L}_{\mathrm{srv}}(W,V) = \mathbb{E}_{\tilde{x}\sim\widetilde{\mathcal{D}}}\mathbb{E}_m \left[\|VW(\tilde{x}\odot m) - \tilde{x}\odot(1-m)\|_2^2 \right]. \tag{31}$$

The practical procedure follows the sequential order of local optimization, model aggregation, and server-side masked training. To analyze the resulting representation without explicitly tracking the intermediate parameter iterates, we use W and V as the common optimization variables and characterize the supervision introduced within one communication round through the following global objective:

$$\mathcal{L}_{\mathrm{Global}}(W,V) = \mathcal{L}_{\mathrm{Fed}}(W,V) + \mathcal{L}_{\mathrm{srv}}(W,V) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\mathrm{loc}}^{(i)}(W,V) + \mathcal{L}_{\mathrm{srv}}(W,V). \tag{32}$$

This one-round surrogate objective provides a tractable formulation for comparing federated learning with and without server-side training on client-generated data.

A.6.2 Technical Assumptions

Based on the formulation above, we next introduce several technical conditions used in the subsequent analysis. These conditions regularize the random sampling process and the geometry of the learned representation, and enable the population-level semantic structure to be related to the empirical encoder obtained under federated optimization.

Assumption A.1. (Independent Sampling.) For each client $i \in [N]$, the observations in the local dataset D_i are independently sampled from the local distribution $\mathcal{D}_i$. Across data realizations, the random variables in the heterogeneous data construction are mutually independent, and the masking variable is sampled independently of the data noise.

Remark A.2. Independent sampling is a standard condition in theoretical analyses of federated and self-supervised representation learning (Wang et al., 2022; Tang et al., 2022; Sun et al., 2024; Chen et al., 2026). It is consistent with the stochastic data model adopted here, where local observations are generated from client-specific distributions and the auxiliary random variables are sampled independently across realizations. Under this condition, the heterogeneity considered in our analysis is attributed to differences among client distributions rather than additional dependence among sampled observations. It also supports the concentration arguments for empirical second-order statistics used in the subsequent proofs.

Assumption A.3. (Non-collapsed Representation.) The learned encoder $W \in \mathbb{R}^{c\times d}$ has full row rank and sufficient capacity to contain the global semantic subspace, i.e.,

$$\operatorname{rank}(W) = c, \qquad N \le c \le d. \tag{33}$$

Remark A.4. Full-row-rank encoders are commonly adopted as a non-degeneracy condition in theoretical analyses of linear representation learning (Liu et al., 2022; Wang et al., 2022; Chen et al., 2026). In our setting, the requirement $c \ge N$ follows naturally from $\dim(\mathcal{S}) = N$, where $\mathcal{S} = \operatorname{span}\{e_1, \dots, e_N\}$. Thus, $N \le c \le d$ ensures that the encoder has sufficient dimensional capacity to represent the semantic subspace without prespecifying which semantic directions are selected by optimization.

Assumption A.5. (Synthetic Semantic Consistency.) Let $P_{\mathcal{S}}$ denote the orthogonal projection onto the global semantic subspace $\mathcal{S}$. The aggregated synthetic second-order statistic $\widetilde{\Sigma}$ satisfies

$$\left\| P_{\mathcal{S}} \left(\widetilde{\Sigma} - \bar{\Sigma} \right) P_{\mathcal{S}} \right\|_2 \le \omega_d S_d, \tag{34}$$

where $\omega_d \to 0$ as $d \to \infty$.

Remark A.6. The above condition is natural for the server-side refinement considered in our framework, where synthetic samples generated from complementary client distributions are aggregated to approximate the global data structure. Related theoretical analyses of federated and distributed optimization commonly impose bounded variance or bounded client dissimilarity conditions to control the discrepancy between local random quantities and their global counterparts (Jing et al., 2024; Tang et al., 2022; Karimireddy et al., 2020). Our assumption follows a similar principle, but is specialized to the representation setting considered here: it only requires the aggregated synthetic second-order statistic to approximate the global statistic within the semantic subspace. The condition $\omega_d \to 0$ further ensures that this approximation error becomes asymptotically negligible relative to the weakest semantic strength S_d.

A.6.3 Theory Results

Lemma A.7. (Global Representability with Generative Refinement). *Consider a federated system with $N = \Theta(d^{1/20})$ clients following the heterogeneous data setup, with $\tau = d^{1/5}$ and $\mu = d^{-1/5}$. Let $\mathcal{S} = \operatorname{span}\{e_1, \dots, e_N\}$ denote the global semantic subspace, whose weakest population semantic strength satisfies*

$$S_d = \frac{1}{4} d^{2/5} + \Theta(d^{-1/20}) + d^{-2/5}. \tag{35}$$

For the global model obtained after federated aggregation and server-side generative refinement, in the non-degenerate regime where $g_d^{\mathrm{Fed}} + \widetilde{b}_{\min} > \sigma\sqrt{S_d/2}$ and under the perturbation condition $\delta_d^{\mathrm{Global}} \le \Delta_{\mathrm{Global}}/2$, with high probability, we have

$$1 - \frac{4(\delta_d^{\mathrm{Global}})^2 \left(h_d^{\mathrm{Fed}} + \widetilde{a}_{\max}\right)^2}{\left(g_d^{\mathrm{Fed}} + \widetilde{b}_{\min} - \sigma\sqrt{S_d/2}\right)^4} \le r_k^{\mathrm{Global}} \le 1 \tag{36}$$

for every $e_k \in \mathcal{S}$. Here, $r_k^{\mathrm{Global}} = \|\Pi_{\mathrm{row}(W_{\mathrm{Global}})} e_k\|_2^2$ measures the representability of semantic direction e_k, h_d^{Fed} and g_d^{Fed} are the federated semantic statistics along the weakest semantic mode, $\widetilde{a}_{\max}$ and $\widetilde{b}_{\min}$ capture the worst-case semantic contributions of the server-generated data, and $\delta_d^{\mathrm{Global}}$ denotes the perturbation around the dominant population semantic representation.

Proof. We begin by characterizing the local optimization on each client. Different from standard MIM, the local objective in our setting jointly optimizes masked reconstruction and complete-data generation through the shared encoder. Therefore, we first derive an equivalent form of the complete local objective and characterize its optimal generative decoder, which is subsequently used to construct the server-side synthetic distribution.

For local optimization. Under the normalized setting $\lambda = \beta = 1$, the local objective on client i is

$$\mathcal{L}_{\mathrm{loc}}^{(i)}(W_i, V_i) = \mathcal{L}_{\mathrm{MIM}}^{(i)}(W_i, V_i) + \mathcal{L}_{\mathrm{VAE}}^{(i)}(W_i, V_i). \tag{37}$$

We first consider the masked reconstruction term. Let

$$A_i = \mathbb{E}_{x\sim\mathcal{D}_i}\mathbb{E}_m\left[x^{\mathrm{vis}}(x^{\mathrm{vis}})^\intercal\right], \qquad B_i = \mathbb{E}_{x\sim\mathcal{D}_i}\mathbb{E}_m\left[x^{\mathrm{vis}}(x^{\mathrm{mask}})^\intercal\right]. \tag{38}$$

Expanding the squared reconstruction loss gives

$$\begin{aligned}\mathcal{L}_{\mathrm{MIM}}^{(i)}(W_i, V_i) &= \mathbb{E}_{x,m}\left[\left\|V_iW_ix^{\mathrm{vis}} - x^{\mathrm{mask}}\right\|_2^2\right] \\ &= \mathbb{E}_{x,m}\left[(x^{\mathrm{vis}})^\intercal W_i^\intercal V_i^\intercal V_iW_ix^{\mathrm{vis}}\right] \\ &\quad - 2\mathbb{E}_{x,m}\left[(x^{\mathrm{mask}})^\intercal V_iW_ix^{\mathrm{vis}}\right] + \mathbb{E}_{x,m}\left[\|x^{\mathrm{mask}}\|_2^2\right].\end{aligned} \tag{39}$$

By the cyclic property of the trace, the above objective can be rewritten as

$$\mathcal{L}_{\mathrm{MIM}}^{(i)}(W_i, V_i) = \mathrm{tr}\left(V_iW_iA_iW_i^\intercal V_i^\intercal\right) - 2\,\mathrm{tr}\left(V_iW_iB_i\right) + C_i^M, \tag{40}$$

where C_i^M collects the terms independent of W_i and V_i. Next, we derive the generative component of the local objective. Define the second-order statistic of the complete local data as

$$\Sigma_i = \mathbb{E}_{x\sim\mathcal{D}_i}\left[xx^\intercal\right]. \tag{41}$$

According to the linear-Gaussian posterior defined in the problem setup, a latent variable can be equivalently expressed through reparameterization as

$$z = W_ix + \sigma\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I_c). \tag{42}$$

Hence, the reconstruction component of the VAE objective satisfies

$$\mathbb{E}_{x\sim\mathcal{D}_i}\mathbb{E}_{z\sim q_{W_i}(z|x)}\left[\|V_iz - x\|_2^2\right] = \mathbb{E}_{x,\epsilon}\left[\|V_iW_ix - x + \sigma V_i\epsilon\|_2^2\right]. \tag{43}$$

Since ϵ is independent of x and satisfies $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}[\epsilon\epsilon^\intercal] = I_c$, the cross term vanishes in expectation. We therefore obtain

$$\mathbb{E}_x\mathbb{E}_{z|x}\left[\|V_iz - x\|_2^2\right] = \mathrm{tr}\left(V_iW_i\Sigma_iW_i^\intercal V_i^\intercal\right) - 2\,\mathrm{tr}\left(V_iW_i\Sigma_i\right) + \sigma^2\|V_i\|_F^2 + C_i^R, \tag{44}$$

where $C_i^R = \mathrm{tr}(\Sigma_i)$ is independent of the model parameters. For the KL regularization, since

$$q_{W_i}(z \mid x) = \mathcal{N}(W_ix, \sigma^2I_c)$$

and $p(z) = \mathcal{N}(0, I_c)$, the KL divergence between the two Gaussian distributions is

$$D_{\mathrm{KL}}\left(q_{W_i}(z \mid x)\,\|\,p(z)\right) = \frac{1}{2}\left[\|W_ix\|_2^2 + c\left(\sigma^2 - 1 - \log\sigma^2\right)\right]. \tag{45}$$

Taking expectation over $x \sim \mathcal{D}_i$ yields

$$\mathbb{E}_{x\sim\mathcal{D}_i}\left[D_{\mathrm{KL}}\left(q_{W_i}(z \mid x)\,\|\,p(z)\right)\right] = \frac{1}{2}\,\mathrm{tr}\left(W_i\Sigma_iW_i^\intercal\right) + C_i^K, \tag{46}$$

where

$$C_i^K = \frac{c}{2}\left(\sigma^2 - 1 - \log\sigma^2\right)$$

is independent of W_i and V_i. Combining Eqs. (40), (44), and (46), define

$$H_i = A_i + \Sigma_i, \qquad G_i = B_i + \Sigma_i. \tag{47}$$

Then, up to an additive constant independent of the model parameters, the complete local objective becomes

$$\mathcal{L}_{\mathrm{loc}}^{(i)}(W_i, V_i) = \mathrm{tr}\left(V_iW_iH_iW_i^\intercal V_i^\intercal\right) - 2\,\mathrm{tr}\left(V_iW_iG_i\right) + \sigma^2\|V_i\|_F^2 + \frac{1}{2}\,\mathrm{tr}\left(W_i\Sigma_iW_i^\intercal\right) + C_i, \tag{48}$$

where $C_i = C_i^M + C_i^R + C_i^K$ collects all constant terms. We next optimize the decoder V_i for a fixed encoder W_i to characterize the client-specific generative model. Since both A_i and Σ_i are symmetric positive semidefinite, H_i is also symmetric positive semidefinite. Taking the derivative of Eq. (48) with respect to V_i gives

$$\frac{\partial \mathcal{L}_{\text{loc}}^{(i)}}{\partial V_i} = 2V_i W_i H_i W_i^\intercal - 2G_i^\intercal W_i^\intercal + 2\sigma^2 V_i. \tag{49}$$

Setting Eq. (49) to zero yields

$$V_i \left(W_i H_i W_i^\intercal + \sigma^2 I_c\right) = G_i^\intercal W_i^\intercal. \tag{50}$$

Since $\sigma^2 > 0$, the matrix $W_i H_i W_i^\intercal + \sigma^2 I_c$ is positive definite and therefore invertible. Hence, for a fixed encoder W_i, the unique optimal decoder is

$$V_i^* = G_i^\intercal W_i^\intercal \left(W_i H_i W_i^\intercal + \sigma^2 I_c\right)^{-1}. \tag{51}$$

The above closed-form solution characterizes the generative decoder learned from the client-specific distribution and will be used to derive the statistics of the synthetic data generated at the server.

For global feature space. We now characterize the global representation resulting from federated aggregation and server-side training. Define the aggregated local statistics

$$\bar{\Sigma} = \frac{1}{N}\sum_{i=1}^N \Sigma_i, \qquad \bar{H} = \frac{1}{N}\sum_{i=1}^N H_i, \qquad \bar{G} = \frac{1}{N}\sum_{i=1}^N G_i. \tag{52}$$

Using the quadratic form of the local objectives, the corresponding federated objective over the common global variables (W, V) is

$$\mathcal{L}_{\text{Fed}}(W,V) = \operatorname{tr}\left(VW\bar{H}W^\intercal V^\intercal\right) - 2\operatorname{tr}\left(VW\bar{G}\right) + \sigma^2\|V\|_F^2 + \frac{1}{2}\operatorname{tr}\left(W\bar{\Sigma}W^\intercal\right) + \bar{C}, \tag{53}$$

where $\bar{C}$ is independent of the model parameters. The locally optimized decoders are further used to generate complete synthetic samples at the server. Since $\tilde{x}_i = V_i^* z_i$ with $z_i \sim \mathcal{N}(0, I_c)$, the second-order statistic of the uniformly mixed synthetic distribution is

$$\widetilde{\Sigma} = \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\tilde{x}_i \tilde{x}_i^\intercal\right] = \frac{1}{N}\sum_{i=1}^N V_i^*(V_i^*)^\intercal. \tag{54}$$

Once generated, $\widetilde{\Sigma}$ is treated as fixed during the subsequent server-side optimization. Let

$$\widetilde{D} = \operatorname{Diag}\left(\operatorname{diag}(\widetilde{\Sigma})\right). \tag{55}$$

Under the same masking rule used for local training, with masking probability p_m, the corresponding server-side masked statistics satisfy

$$\widetilde{A} = (1-p_m)^2\widetilde{\Sigma} + p_m(1-p_m)\widetilde{D}, \qquad \widetilde{B} = p_m(1-p_m)\left(\widetilde{\Sigma} - \widetilde{D}\right). \tag{56}$$

The server-side reconstruction loss therefore has the same quadratic form as Eq. (40). Define

$$H_{\text{Global}} = \bar{H} + \widetilde{A}, \qquad G_{\text{Global}} = \bar{G} + \widetilde{B}. \tag{57}$$

Then the complete global objective $\mathcal{L}_{\text{Global}} = \mathcal{L}_{\text{Fed}} + \mathcal{L}_{\text{srv}}$ can be expressed as

$$\begin{aligned}\mathcal{L}_{\text{Global}}(W,V) = {} & \operatorname{tr}\left(VWH_{\text{Global}}W^\intercal V^\intercal\right) - 2\operatorname{tr}\left(VWG_{\text{Global}}\right) \\ & + \sigma^2\|V\|_F^2 + \frac{1}{2}\operatorname{tr}\left(W\bar{\Sigma}W^\intercal\right) + C_{\text{Global}}.\end{aligned} \tag{58}$$

By similar arguments to the local decoder optimization, for any fixed W, the unique optimal global decoder is

$$V_{\text{Global}}^* = G_{\text{Global}}^\intercal W^\intercal \left(WH_{\text{Global}}W^\intercal + \sigma^2 I_c\right)^{-1}. \tag{59}$$

Substituting Eq. (59) into Eq. (58) yields

$$\begin{aligned}\mathcal{L}_{\text{Global}}(W) = {} & -\operatorname{tr}\left[G_{\text{Global}}^\intercal W^\intercal \left(WH_{\text{Global}}W^\intercal + \sigma^2 I_c\right)^{-1} WG_{\text{Global}}\right] \\ & + \frac{1}{2}\operatorname{tr}\left(W\bar{\Sigma}W^\intercal\right) + C_{\text{Global}}.\end{aligned} \tag{60}$$

We next characterize the representation induced by Eq. (60). Let v_k denote a common population eigendirection of the dominant aggregated local statistics, such that

$$\bar{\Sigma} v_k = s_k v_k, \qquad \bar{H} v_k = h_k^{\mathrm{Fed}} v_k, \qquad \bar{G} v_k = g_k^{\mathrm{Fed}} v_k. \tag{61}$$

For the server-generated component, define its directional contributions as

$$\widetilde{a}_k = v_k^{\intercal} \widetilde{A} v_k, \qquad \widetilde{b}_k = v_k^{\intercal} \widetilde{B} v_k. \tag{62}$$

The remaining off-direction components are incorporated into the perturbation term introduced later. For an encoder component $w_k = \sqrt{t_k} v_k^{\intercal}$ with $t_k \geq 0$, we have

$$\|G_{\mathrm{Global}} v_k\|_2^2 \geq \left(v_k^{\intercal} G_{\mathrm{Global}} v_k\right)^2 = \left(g_k^{\mathrm{Fed}} + \widetilde{b}_k\right)^2. \tag{63}$$

Accordingly, the directional contribution to the reduced objective is bounded by

$$\ell_k^{\mathrm{Global}}(t_k) \leq -\frac{t_k \left(g_k^{\mathrm{Fed}} + \widetilde{b}_k\right)^2}{t_k \left(h_k^{\mathrm{Fed}} + \widetilde{a}_k\right) + \sigma^2} + \frac{1}{2} t_k s_k. \tag{64}$$

Minimizing the right-hand side of Eq. (64) over $t_k \geq 0$ yields the following lower bound on the effective representation score:

$$\rho_k^{\mathrm{Global}} \geq \frac{\left(\left|g_k^{\mathrm{Fed}} + \widetilde{b}_k\right| - \sigma\sqrt{s_k/2}\right)_+^2}{h_k^{\mathrm{Fed}} + \widetilde{a}_k}, \tag{65}$$

where $(a)_+ = \max\{a, 0\}$. Then, we specify this score under the non-IID data construction. Consider

$$N = \Theta(d^{1/20}), \qquad \tau = d^{1/5}, \qquad \mu = d^{-1/5}. \tag{66}$$

Let $\mathcal{S} = \mathrm{span}\{e_1, \ldots, e_N\}$ denote the global semantic subspace. Under Assumption A.1 and the symmetric dominant-class construction, the dominant population component of the second-order statistic restricted to $\mathcal{S}$ takes the form

$$\bar{\Sigma}_{\mathcal{S}} = (\chi_d - \psi_d) I_N + \psi_d \mathbf{1}\mathbf{1}^{\intercal}, \tag{67}$$

where χ_d and ψ_d denote the diagonal and off-diagonal strengths induced by the dominant client-specific semantic components, respectively. The lower-order contribution of the sparsely observed non-dominant samples is absorbed into the perturbation term introduced later. Specifically,

$$\chi_d = \frac{1}{N} + \frac{N-1}{2N}\tau^2 + \mu^2, \qquad \psi_d = \frac{N-2}{4N}\tau^2. \tag{68}$$

Since a matrix of the form $(\chi_d - \psi_d) I_N + \psi_d \mathbf{1}\mathbf{1}^{\intercal}$ has eigenvalue $\chi_d + (N-1)\psi_d$ along $\mathbf{1}$ and eigenvalue $\chi_d - \psi_d$ on its orthogonal complement, its smallest semantic eigenvalue is

$$\begin{aligned} S_d := \lambda_{\min}\left(\bar{\Sigma}_{\mathcal{S}}\right) &= \chi_d - \psi_d \\ &= \frac{1}{N} + \frac{\tau^2}{4} + \mu^2 \\ &= \frac{1}{4} d^{2/5} + \Theta(d^{-1/20}) + d^{-2/5}. \end{aligned} \tag{69}$$

As the diagonal part of $\bar{\Sigma}_{\mathcal{S}}$ equals $\chi_d I_N$, define the corresponding federated semantic scalars along the weakest semantic mode as

$$\begin{aligned} h_d^{\mathrm{Fed}} &= \left[1 + (1-p_m)^2\right] S_d + p_m(1-p_m)\chi_d, \\ g_d^{\mathrm{Fed}} &= \left[1 + p_m(1-p_m)\right] S_d - p_m(1-p_m)\chi_d. \end{aligned} \tag{70}$$

For the generated-data contribution, define the worst-case semantic directional statistics

$$\widetilde{a}_{\max} = \max_{k \in [N]} \widetilde{a}_k, \qquad \widetilde{b}_{\min} = \min_{k \in [N]} \widetilde{b}_k. \tag{71}$$

The weakest semantic mode S_d determines the asymptotic bottleneck of the population semantic representation. Therefore, provided that

$$g_d^{\mathrm{Fed}} + \widetilde{b}_{\min} > \sigma\sqrt{S_d/2}, \tag{72}$$

the smallest semantic representation score satisfies

$$\min_{k\in[N]} \rho_{k,\mathrm{sem}}^{\mathrm{Global}} \geq \frac{\left(g_d^{\mathrm{Fed}} + \widetilde{b}_{\min} - \sigma\sqrt{S_d/2}\right)^2}{h_d^{\mathrm{Fed}} + \widetilde{a}_{\max}}. \tag{73}$$

Under the dominant population block structure above, the semantic directions decouple in the reduced objective, and directions with larger effective scores are preferentially represented. Accordingly, define

$$X_{\mathrm{Global}}^{\mathrm{rep}} = \sum_{k=1}^{N} \rho_{k,\mathrm{sem}}^{\mathrm{Global}} v_k v_k^{\intercal} \tag{74}$$

as the dominant population semantic representation operator. Assumption A.3 ensures that the encoder has sufficient non-degenerate capacity to represent the semantic subspace, allowing its learned row space to be characterized through the dominant eigenspace of the corresponding representation operator. Together with the independent sampling structure in Assumption A.1, standard matrix concentration arguments (Tropp, 2012) and the lower-order residual components, with high probability, we have

$$\left\| \widehat{X}_{\mathrm{Global}}^{\mathrm{rep}} - X_{\mathrm{Global}}^{\mathrm{rep}} \right\|_2 \leq \delta_d^{\mathrm{Global}}, \tag{75}$$

where $\delta_d^{\mathrm{Global}}$ collects the resulting perturbation around the dominant population semantic representation operator. Since $X_{\mathrm{Global}}^{\mathrm{rep}}$ is supported on $\mathcal{S}$, its semantic separation satisfies

$$\Delta_{\mathrm{Global}} \geq \frac{\left(g_d^{\mathrm{Fed}} + \widetilde{b}_{\min} - \sigma\sqrt{S_d/2}\right)^2}{h_d^{\mathrm{Fed}} + \widetilde{a}_{\max}}. \tag{76}$$

If $\delta_d^{\mathrm{Global}} \leq \Delta_{\mathrm{Global}}/2$, then by the Davis-Kahan eigenspace perturbation theorem (Davis & Kahan, 1970), for every $e_k \in \mathcal{S}$, there exists

$$\begin{aligned} r_k^{\mathrm{Global}} &= \left\| \Pi_{\mathrm{row}(W_{\mathrm{Global}})} e_k \right\|_2^2 \\ &\geq 1 - \frac{4(\delta_d^{\mathrm{Global}})^2 \left(h_d^{\mathrm{Fed}} + \widetilde{a}_{\max}\right)^2}{\left(g_d^{\mathrm{Fed}} + \widetilde{b}_{\min} - \sigma\sqrt{S_d/2}\right)^4}. \end{aligned} \tag{77}$$

This completes the proof.

□

Corollary A.8. (Global Representability of Federated MIM). *Under the same heterogeneous data setup and scaling conditions as in Lemma A.7, consider standard federated MIM without local generative learning or server-side generative refinement. For any fixed masking probability $p_m \in (0,1)$, under the perturbation condition, $\delta_d^{\mathrm{MIM}} \leq \Delta_{\mathrm{FedMIM}}/2$, with high probability, for every $e_k \in \mathcal{S}$, it satisfies*

$$1 - \frac{4(\delta_d^{\mathrm{MIM}})^2}{\left[\frac{p_m^2(1-p_m)}{4(1+p_m)} d^{2/5} - O(d^{7/20})\right]^2} \leq r_k^{\mathrm{FedMIM}} \leq 1. \tag{78}$$

Here, $r_k^{\mathrm{FedMIM}} = \|\Pi_{\mathrm{row}(W_{\mathrm{FedMIM}})} e_k\|_2^2$ measures the representability of semantic direction e_k learned by standard federated MIM, and δ_d^{MIM} denotes the perturbation around its dominant population semantic representation.

Proof. Standard federated MIM optimizes only the masked reconstruction objective on each client, without either local generative learning or server-side synthetic refinement. Therefore, using Eq. (40) and averaging over all clients, its global objective can be written as

$$\mathcal{L}_{\mathrm{FedMIM}}(W,V) = \mathrm{tr}\left(VW\bar{A}W^{\intercal}V^{\intercal}\right) - 2\,\mathrm{tr}\left(VW\bar{B}\right) + C_{\mathrm{MIM}}, \tag{79}$$

where

$$\bar{A} = \frac{1}{N}\sum_{i=1}^{N} A_i, \qquad \bar{B} = \frac{1}{N}\sum_{i=1}^{N} B_i. \tag{80}$$

Different from Lemma A.7, Eq. (79) contains neither the VAE-induced complete-data statistics nor the server-generated reconstruction term. Under the same masking rule, let

$$\bar{D} = \mathrm{Diag}\left(\mathrm{diag}(\bar{\Sigma})\right). \tag{81}$$

Then the aggregated masked statistics satisfy

$$\bar{A} = (1-p_m)^2\bar{\Sigma} + p_m(1-p_m)\bar{D}, \qquad \bar{B} = p_m(1-p_m)\left(\bar{\Sigma} - \bar{D}\right). \tag{82}$$

By similar arguments to the decoder elimination in Lemma A.7, minimizing Eq. (79) over V gives

$$V^*_{\mathrm{FedMIM}} = \bar{B}^\intercal W^\intercal \left(W\bar{A}W^\intercal\right)^\dagger, \tag{83}$$

where $(\cdot)^\dagger$ denotes the Moore–Penrose pseudoinverse. Substituting Eq. (83) back into the objective yields

$$\mathcal{L}_{\mathrm{FedMIM}}(W) = -\operatorname{tr}\left[\bar{B}^\intercal W^\intercal \left(W\bar{A}W^\intercal\right)^\dagger W\bar{B}\right] + C_{\mathrm{MIM}}. \tag{84}$$

We next specialize the above objective to the dominant semantic subspace $\mathcal{S}$. Following Assumption A.1 and the same population construction as in Lemma A.7, we have

$$\bar{\Sigma}_{\mathcal{S}} = (\chi_d - \psi_d)I_N + \psi_d \mathbf{1}\mathbf{1}^\intercal, \qquad \bar{D}_{\mathcal{S}} = \chi_d I_N, \tag{85}$$

$\bar{\Sigma}$, $\bar{A}$, and $\bar{B}$ share the same population semantic eigendirections. Thus, following the same directional reduction as in Lemma A.7, the effective representation score along a semantic direction v_k is

$$\rho_k^{\mathrm{MIM}} = \frac{\left(g_k^{\mathrm{MIM}}\right)^2}{h_k^{\mathrm{MIM}}}, \tag{86}$$

where

$$h_k^{\mathrm{MIM}} = v_k^\intercal \bar{A} v_k, \qquad g_k^{\mathrm{MIM}} = v_k^\intercal \bar{B} v_k. \tag{87}$$

For the weakest semantic mode with strength S_d, Eq. (82) gives

$$\begin{aligned} h_d^{\mathrm{MIM}} &= (1-p_m)^2 S_d + p_m(1-p_m)\chi_d, \\ g_d^{\mathrm{MIM}} &= p_m(1-p_m)\left(S_d - \chi_d\right) \\ &= -p_m(1-p_m)\psi_d. \end{aligned} \tag{88}$$

We now substitute the scaling

$$N = \Theta(d^{1/20}), \qquad \tau = d^{1/5}, \qquad \mu = d^{-1/5}. \tag{89}$$

From the semantic statistics derived in Lemma A.7, we have

$$\begin{aligned} S_d &= \frac{1}{4}d^{2/5} + \Theta(d^{-1/20}) + d^{-2/5}, \\ \chi_d &= \frac{1}{2}d^{2/5} - \Theta(d^{7/20}) + O(d^{-1/20}), \\ \psi_d &= \frac{1}{4}d^{2/5} - \Theta(d^{7/20}). \end{aligned} \tag{90}$$

Substituting Eq. (90) into Eq. (88) gives

$$\begin{aligned} h_d^{\mathrm{MIM}} &= \frac{1-p_m^2}{4}d^{2/5} - O(d^{7/20}), \\ \left|g_d^{\mathrm{MIM}}\right| &= \frac{p_m(1-p_m)}{4}d^{2/5} - O(d^{7/20}). \end{aligned} \tag{91}$$

Hence, using Eq. (86), the weakest semantic representation score satisfies

$$\begin{aligned} \min_{k\in[N]} \rho_{k,\mathrm{sem}}^{\mathrm{MIM}} &\geq \frac{\left(g_d^{\mathrm{MIM}}\right)^2}{h_d^{\mathrm{MIM}}} \\ &\geq \frac{p_m^2(1-p_m)}{4(1+p_m)}d^{2/5} - O(d^{7/20}). \end{aligned} \tag{92}$$

By the same population-operator construction used in Lemma A.7, let $X^{\mathrm{rep}}_{\mathrm{FedMIM}}$ denote the corresponding population semantic representation operator. The non-collapse condition in Assumption A.3 guarantees sufficient encoder capacity for characterizing the learned row space through the dominant eigenspace of this operator. Then, combining with AssumptionA.1 and the empirical matrix concentration arguments (Tropp, 2012), with high probability, there exists

$$\left\|\widehat{X}^{\mathrm{rep}}_{\mathrm{FedMIM}} - X^{\mathrm{rep}}_{\mathrm{FedMIM}}\right\|_2 \le \delta_d^{\mathrm{MIM}}. \tag{93}$$

Accordingly, the population semantic separation satisfies

$$\Delta_{\mathrm{FedMIM}} \ge \frac{p_m^2(1-p_m)}{4(1+p_m)} d^{2/5} - O(d^{7/20}). \tag{94}$$

Under the corresponding perturbation condition $\delta_d^{\mathrm{MIM}} \le \Delta_{\mathrm{FedMIM}}/2$, by the same Davis-Kahan argument (Davis & Kahan, 1970), for every $e_k \in \mathcal{S}$, we have

$$\begin{aligned} r_k^{\mathrm{FedMIM}} &= \left\|\Pi_{\mathrm{row}(W_{\mathrm{FedMIM}})} e_k\right\|_2^2 \\ &\ge 1 - \frac{4(\delta_d^{\mathrm{MIM}})^2}{\Delta^2_{\mathrm{FedMIM}}} \\ &\ge 1 - \frac{4(\delta_d^{\mathrm{MIM}})^2}{\left[\frac{p_m^2(1-p_m)}{4(1+p_m)} d^{2/5} - O(d^{7/20})\right]^2}. \end{aligned} \tag{95}$$

Since $r_k^{\mathrm{FedMIM}} \le 1$ by definition, the proof has been completed. □

Theorem A.9. (Corresponding to Theorem 3.1 in manuscript). *Under the same heterogeneous data setup and scaling conditions as in Lemma A.7 and Corollary A.8, let* $\delta_d = \max\left\{\delta_d^{\mathrm{MIM}}, \delta_d^{\mathrm{Gen}}, \delta_d^{\mathrm{Global}}\right\}$ *be a common upper bound on the corresponding population-to-empirical representation perturbations. For any fixed* $p_m \in (0,1)$ *and sufficiently large* d*, the following results hold for comparisons with the representation learned by federated MIM.*

1. *Let* $r_k^{\mathrm{Gen}} = \|\Pi_{\mathrm{row}(W_{\mathrm{Gen}})} e_k\|_2^2$ *denote the representability obtained after local generative learning but before server-side refinement. Then, for every* $e_k \in \mathcal{S}$*,*

$$r_k^{\mathrm{Gen}} \ge 1 - \frac{4\delta_d^2}{\left[C_{\mathrm{Gen}}(p_m) d^{2/5} + o(d^{2/5})\right]^2} > 1 - \frac{4\delta_d^2}{\left[C_{\mathrm{MIM}}(p_m) d^{2/5} + o(d^{2/5})\right]^2}, \tag{96}$$

   where $C_{\mathrm{Gen}}(p_m) = \frac{[1-p_m(1-p_m)]^2}{4(2-p_m^2)}$ *and* $C_{\mathrm{MIM}}(p_m) = \frac{p_m^2(1-p_m)}{4(1+p_m)}$.

2. *When the aggregated synthetic distribution preserves the global semantic statistics consistently, let* $r_k^{\mathrm{Global}} = \|\Pi_{\mathrm{row}(W_{\mathrm{Global}})} e_k\|_2^2$ *denote the representability after server-side generative refinement. Then, for every* $e_k \in \mathcal{S}$*,*

$$r_k^{\mathrm{Global}} \ge 1 - \frac{4\delta_d^2}{\left[C_{\mathrm{Global}}(p_m) d^{2/5} + o(d^{2/5})\right]^2} > 1 - \frac{4\delta_d^2}{\left[C_{\mathrm{MIM}}(p_m) d^{2/5} + o(d^{2/5})\right]^2}, \tag{97}$$

   where $C_{\mathrm{Global}}(p_m) = \frac{[1-2p_m(1-p_m)]^2}{4(3-2p_m^2)}$.

Proof. For a unified comparison, let $X^{\mathrm{rep}}_{\star}$ and $\widehat{X}^{\mathrm{rep}}_{\star}$ denote the population and empirical representation operators, respectively, for $\star \in \{\mathrm{MIM}, \mathrm{Gen}, \mathrm{Global}\}$. Let their perturbation levels satisfy

$$\left\|\widehat{X}^{\mathrm{rep}}_{\star} - X^{\mathrm{rep}}_{\star}\right\|_2 \le \delta_d^{\star}, \qquad \star \in \{\mathrm{MIM}, \mathrm{Gen}, \mathrm{Global}\}. \tag{98}$$

We define the common perturbation upper bound as

$$\delta_d := \max\left\{\delta_d^{\mathrm{MIM}}, \delta_d^{\mathrm{Gen}}, \delta_d^{\mathrm{Global}}\right\}. \tag{99}$$

Hence, $\delta_d^\star \le \delta_d$ for all three representations. Replacing each individual perturbation by δ_d therefore gives a common conservative lower bound and allows the three representability guarantees to be compared under the same perturbation scale. Then, let

$$q_m := p_m(1-p_m).$$

We prove the theorem in two parts.

Part I: Improvement from local generative learning. For standard federated MIM, following Corollary A.8, its weakest semantic statistics satisfy

$$\begin{aligned} h_d^{\mathrm{MIM}} &= \frac{1-p_m^2}{4} d^{2/5} - O(d^{7/20}), \\ \left|g_d^{\mathrm{MIM}}\right| &= \frac{p_m(1-p_m)}{4} d^{2/5} - O(d^{7/20}). \end{aligned} \tag{100}$$

Accordingly, Corollary A.8 gives

$$\Delta_{\mathrm{FedMIM}} \ge \frac{p_m^2(1-p_m)}{4(1+p_m)} d^{2/5} - O(d^{7/20}). \tag{101}$$

Since $d^{7/20} = o(d^{2/5})$, Eq. (101) can be equivalently expressed in leading-order form as

$$\Delta_{\mathrm{FedMIM}} \ge C_{\mathrm{MIM}}(p_m) d^{2/5} + o(d^{2/5}), \tag{102}$$

where

$$C_{\mathrm{MIM}}(p_m) = \frac{p_m^2(1-p_m)}{4(1+p_m)}. \tag{103}$$

From Corollary A.8, the original FedMIM representability guarantee contains $(\delta_d^{\mathrm{MIM}})^2$ in its perturbation term. Since $\delta_d^{\mathrm{MIM}} \le \delta_d$, we have

$$\begin{aligned} r_k^{\mathrm{FedMIM}} &\ge 1 - \frac{4(\delta_d^{\mathrm{MIM}})^2}{\left[C_{\mathrm{MIM}}(p_m) d^{2/5} + o(d^{2/5})\right]^2} \\ &\ge 1 - \frac{4\delta_d^2}{\left[C_{\mathrm{MIM}}(p_m) d^{2/5} + o(d^{2/5})\right]^2}. \end{aligned} \tag{104}$$

We next consider the model with local generative learning but without server-side synthetic refinement. Since the generative objective contributes an additional complete-data statistic, its weakest-mode statistics satisfy

$$h_d^{\mathrm{Fed}} = S_d + h_d^{\mathrm{MIM}}, \qquad g_d^{\mathrm{Fed}} = S_d + g_d^{\mathrm{MIM}}. \tag{105}$$

Using the semantic scaling established in Lemma A.7 and Corollary A.8, we therefore obtain

$$h_d^{\mathrm{Fed}} = (2-p_m^2) S_d + o(S_d), \qquad g_d^{\mathrm{Fed}} = (1-q_m) S_d + o(S_d). \tag{106}$$

Since $\sigma\sqrt{S_d/2} = o(S_d)$, the corresponding semantic separation becomes

$$\Delta_{\mathrm{Gen}} \ge C_{\mathrm{Gen}}(p_m) d^{2/5} + o(d^{2/5}), \tag{107}$$

where

$$C_{\mathrm{Gen}}(p_m) = \frac{[1-p_m(1-p_m)]^2}{4(2-p_m^2)}. \tag{108}$$

By the same population-operator and Davis-Kahan arguments (Davis & Kahan, 1970), the representability after local generative learning satisfies

$$\begin{aligned} r_k^{\mathrm{Gen}} &\ge 1 - \frac{4(\delta_d^{\mathrm{Gen}})^2}{\left[C_{\mathrm{Gen}}(p_m) d^{2/5} + o(d^{2/5})\right]^2} \\ &\ge 1 - \frac{4\delta_d^2}{\left[C_{\mathrm{Gen}}(p_m) d^{2/5} + o(d^{2/5})\right]^2}. \end{aligned} \tag{109}$$

Furthermore,

$$C_{\mathrm{Gen}}(p_m) - C_{\mathrm{MIM}}(p_m) = \frac{3p_m^3 - p_m^2 - p_m + 1}{4(1+p_m)(2-p_m^2)} > 0 \tag{110}$$

for every $p_m \in (0,1)$. Hence, for sufficiently large d, the dominant $d^{2/5}$ term determines the denominator of the corresponding asymptotic representability bounds. Under the common perturbation upper bound δ_d, we therefore obtain

$$1 - \frac{4\delta_d^2}{\left[C_{\mathrm{Gen}}(p_m)d^{2/5} + o(d^{2/5})\right]^2} > 1 - \frac{4\delta_d^2}{\left[C_{\mathrm{MIM}}(p_m)d^{2/5} + o(d^{2/5})\right]^2}. \tag{111}$$

Therefore, local generative learning alone yields a strictly tighter asymptotic representability lower bound than standard federated MIM.

Part II: Improvement of the complete generative framework. Under Assumption A.5, we have

$$\left\| P_{\mathcal{S}}(\widetilde{\Sigma} - \bar{\Sigma})P_{\mathcal{S}} \right\|_2 \le \omega_d S_d, \qquad \omega_d \to 0. \tag{112}$$

Thus,

$$P_{\mathcal{S}}(\widetilde{\Sigma} - \bar{\Sigma})P_{\mathcal{S}} = o_{\mathrm{op}}(S_d). \tag{113}$$

Since $\mathcal{S}$ is a coordinate semantic subspace, the same order holds for the corresponding diagonal statistics. By the linear masking transformation, we consequently obtain

$$P_{\mathcal{S}}(\widetilde{A} - \bar{A})P_{\mathcal{S}} = o_{\mathrm{op}}(S_d), \qquad P_{\mathcal{S}}(\widetilde{B} - \bar{B})P_{\mathcal{S}} = o_{\mathrm{op}}(S_d). \tag{114}$$

Using $\bar{H} = \bar{\Sigma} + \bar{A}$ and $\bar{G} = \bar{\Sigma} + \bar{B}$, the complete semantic statistics satisfy

$$\begin{aligned} P_{\mathcal{S}} H_{\mathrm{Global}} P_{\mathcal{S}} &= P_{\mathcal{S}}(\bar{\Sigma} + 2\bar{A})P_{\mathcal{S}} + o_{\mathrm{op}}(S_d), \\ P_{\mathcal{S}} G_{\mathrm{Global}} P_{\mathcal{S}} &= P_{\mathcal{S}}(\bar{\Sigma} + 2\bar{B})P_{\mathcal{S}} + o_{\mathrm{op}}(S_d). \end{aligned} \tag{115}$$

Therefore, along the weakest semantic mode,

$$\begin{aligned} h_d^{\mathrm{Global}} &= (3 - 2p_m^2)S_d + o(S_d), \\ g_d^{\mathrm{Global}} &= [1 - 2p_m(1-p_m)]S_d + o(S_d). \end{aligned} \tag{116}$$

Since $p_m(1-p_m) \le 1/4$ and $\sigma\sqrt{S_d/2} = o(S_d)$, the resulting semantic separation satisfies

$$\Delta_{\mathrm{Global}} \ge C_{\mathrm{Global}}(p_m)d^{2/5} + o(d^{2/5}), \tag{117}$$

where

$$C_{\mathrm{Global}}(p_m) = \frac{[1 - 2p_m(1-p_m)]^2}{4(3 - 2p_m^2)}. \tag{118}$$

The original complete-model representability bound contains the corresponding perturbation $\delta_d^{\mathrm{Global}}$. Since $\delta_d^{\mathrm{Global}} \le \delta_d$, we obtain

$$\begin{aligned} r_k^{\mathrm{Global}} &\ge 1 - \frac{4(\delta_d^{\mathrm{Global}})^2}{\left[C_{\mathrm{Global}}(p_m)d^{2/5} + o(d^{2/5})\right]^2} \\ &\ge 1 - \frac{4\delta_d^2}{\left[C_{\mathrm{Global}}(p_m)d^{2/5} + o(d^{2/5})\right]^2}. \end{aligned} \tag{119}$$

Moreover,

$$C_{\mathrm{Global}}(p_m) - C_{\mathrm{MIM}}(p_m) = \frac{2p_m^5 - 2p_m^4 + 3p_m^3 + p_m^2 - 3p_m + 1}{4(1+p_m)(3-2p_m^2)} > 0 \tag{120}$$

for every $p_m \in (0,1)$. Hence, for sufficiently large d, the dominant $d^{2/5}$ term again determines the denominator of the corresponding asymptotic representability bounds. Under the same common perturbation upper bound δ_d, this yields

$$1 - \frac{4\delta_d^2}{\left[C_{\mathrm{Global}}(p_m)d^{2/5} + o(d^{2/5})\right]^2} > 1 - \frac{4\delta_d^2}{\left[C_{\mathrm{MIM}}(p_m)d^{2/5} + o(d^{2/5})\right]^2}. \tag{121}$$

Thus, the complete model with server-side generative refinement admits a strictly tighter asymptotic representability lower bound than standard federated MIM. This completes the proof. □