SEM2PIX: PRIVACY LEAKAGE FROM GRADIENTS VIA SEMANTIC-TO-PIXEL FINE-GRAINED GUIDANCE DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning (FL) enables collaborative model training among multiple clients without requiring them to expose their raw images. Several studies have introduced attacks known as gradient inversion attacks (GIAs), which exploit the gradients shared during training to reconstruct clients' raw images. Recent work has shown that training-free diffusion models can transform Gaussian noise into recovered private images under guidance from shared gradients without requiring knowledge of the clients' data distribution. Our observations suggest that (i) guidance should constrain semantic information, such as global shape and color distribution, and (ii) additional guidance is needed to recover fine-grained details, which are particularly important in sensitive domains such as healthcare. In this paper, we present Semantic-to-Pixel Fine-Grained Guidence (SEM2PIX), which leverages vision–language models (VLMs) to enhance guidance for diffusion-based GIAs. First, we introduce VLM guidance for reverse diffusion to constrain the reconstructed image to be semantically meaningful. Second, we enhance the recovery of fine-grained details through stage-wise gradient guidance, which dynamically matches the layer-wise gradients induced by the predicted clean image with the corresponding shared client gradients. With these two steps, SEM2PIX can recover high-resolution images with high fidelity in practical federated learning scenarios involving defense mechanisms, larger batch sizes, and multiple local updates.

1 INTRODUCTION

1.1 BACKGROUND

Federated learning (FL) allows multiple clients to train a shared model without sending their raw data (Kairouz & McMahan, 2021). However, the gradients shared with the server can reveal private information. Gradient inversion attacks (GIAs) exploit these gradients to reconstruct client's training samples (Zhu et al., 2019; Wu et al., 2023). This threat is especially serious in healthcare imaging, where confidential patient information may be exposed (Raswa et al., 2025; Tsai et al., 2026).

Existing GIAs include pixel-based attacks (Zhu et al., 2019; Wu et al., 2023; Ye et al., 2024; Wan et al., 2026; Fan et al., 2025b; Gong et al., 2026), GAN-based attacks (Li et al., 2022; Fang et al., 2023; Jeon et al., 2021; Xiang et al., 2026), and diffusion-based methods (Meng et al., 2024; Gu et al., 2024; Meng et al., 2026; Carletti et al., 2026). Pixel-based attacks directly optimize an image to match the shared client gradient, but they are expensive at high resolutions and with perturbed gradients. GAN-based attacks search within the output space of a pretrained generator, but their performance depends on how closely the auxiliary training data distribution matches the client data distribution. Diffusion-based methods instead guide a training-free diffusion model using gradients (Guo et al., 2026; 2024). These methods can reconstruct high-quality images without prior knowledge of the client data distribution. This property is useful when the server does not know the private client distribution. More detailed discussions of related works can be found in Sec. A.1 of Appendix.

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.

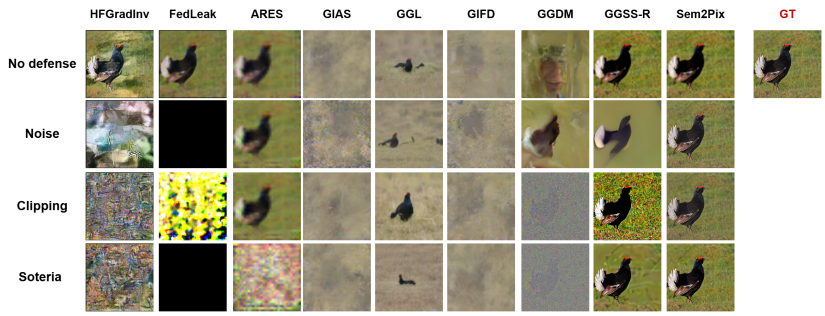


Figure 1: Reconstruction results for representative GIAs without a defense and against noise, gradient clipping, and Soteria. The defenses reduce reconstruction quality by changing the information retained in the perturbed client gradient.

1.2 Observation and Contribution

We study GIAs under three practical FL defenses, such as noise (Wei et al., 2020), gradient clipping, and Soteria (Sun et al., 2021). We refer to the resulting shared gradient as the *perturbed client gradient*. These defenses perturb the shared client gradient and reduce the information available to the attacker. As shown in Figure 1, representative GIAs somewhat fail to recover class-compatible global structure and fine-grained details. Our observations identify three challenges:

- **Label leakage.** GIAs use inferred labels to guide generative priors toward the target class (Li et al., 2022; Fang et al., 2023; Jeon et al., 2021). They infer the private label by examining each row of the final fully connected (FC) layer and selecting the class corresponding to a row whose values are all negative (Zhao et al., 2020; Fang et al., 2023). However, gradient perturbations can turn some negative gradient values positive, making label inference unreliable (Hu et al., 2026).
- **Semantic guidance.** Existing diffusion-based GIAs primarily rely on gradient matching to guide reverse sampling (Meng et al., 2024; Gu et al., 2024; Meng et al., 2026). While gradient matching steers the reverse process toward consistency with the observed client gradient, it does not explicitly incorporate semantic information about the target class. As a result, the reverse trajectory may fail to recover semantic information about the target class, such as its global shape and color distribution.
- **Fine-grained refinement.** Existing diffusion-based GIAs lack explicit pixel-level guidance for recovering fine-grained details (Gu et al., 2024). This limitation is particularly important in healthcare imaging, where the images contain privacy-sensitive clinical information (Raswa et al., 2025; Tsai et al., 2026).

To address these challenges, we present Semantic-to-Pixel Fine-Grained Guidance (Sem2Pix), a training-free diffusion-based GIA that leverages vision–language models (VLMs) (Radford et al., 2021; Yang et al., 2026; Chen et al., 2025) for semantic guidance. To our knowledge, this is the first diffusion-based GIA utilizing VLM guidance under an honest-but-curious server without modifying the global model or FL objective. First, Sem2Pix constructs target text prompts from the inferred private labels and uses VLM guidance to align the reverse trajectory with the semantic information encoded in these prompts. Second, it introduces stage-wise gradient guidance to dynamically match the layer-wise gradients induced by the predicted clean image with the corresponding shared client gradients, thereby progressively recovering fine-grained details. Our contributions are summarized as follows:

- **Magnitude-aware label leakage.** To address the limitation of label leakage in the 1st observation, we propose magnitude-aware label leakage, which requires only the presence of negative gradient values and measures their magnitudes within each row of the FC layer gradient. We then rank the rows by these magnitudes and infer the private labels as the classes corresponding to the highest-ranked rows.

- **Vision–language semantic guidance.** To address the lack of explicit semantic guidance in the 2nd observation, we construct text prompts from the inferred labels and use a pretrained vision–language model to guide the reverse process of a training-free diffusion model (Guo et al., 2026; 2024; Kim et al., 2026). This training-free guidance aligns the reverse trajectory with the semantic information encoded in the target text prompt.
- **Stage-wise gradient guidance.** To address the lack of fine-grained details in the 3rd observation, we introduce stage-wise gradient guidance for the remaining reverse steps. Inspired by HFGradInv (Ye et al., 2024), it dynamically matches the layer-wise gradients induced by the predicted clean image with the corresponding shared client gradients. The matching begins with the final fully connected layer and progressively incorporates earlier feature-extraction layers to recover fine-grained details.

We conduct extensive experiments on high-resolution natural-image datasets, including CelebA (Karras et al., 2018), LSUN (Yu et al., 2015), and ImageNet (Deng et al., 2009). We also evaluate SEM2PIX on Camelyon17 (Bandi et al., 2018) to assess its effectiveness in the privacy-sensitive healthcare imaging domain. Our evaluation considers diverse data distributions and more practical federated learning (FL) settings to assess performance (Carletti et al., 2025).

2 PRELIMINARIES

2.1 THREAT MODEL

We assume an honest-but-curious server that follows the FL training protocol but attempts to reconstruct private client data from the shared client gradient (Zhu et al., 2019). This assumption reflects practical FL settings (Carletti et al., 2025) more closely than a malicious-server model that changes the global model to amplify information leakage (Shan et al., 2025). In the FL setting, we denote the set of participating clients by $\mathcal{C}$ and consider a target client $c \in \mathcal{C}$ with a local batch of size B. The batch size satisfies $1 \leq B \leq K$, where K denotes the total number of classes. Each input image $x \in \mathcal{X}$ has a class label $k \in \{1, \ldots, K\}$, where $\mathcal{X}$ denotes the image domain.

Server Capabilities. Following the standard FL protocol (McMahan et al., 2017), the server has access to the current global model and the shared client gradient. The server knows the optimizer, learning rate, client batch size B, and defense mechanism (Liu et al., 2025; Hu et al., 2026). The server also knows the classifier label vocabulary and the mapping between output indices and class names. Moreover, the server incorporates gradient-matching term specific to the fully connected (FC) layer to leak the client's private label (Zhao et al., 2020; Fang et al., 2023; Ye et al., 2024). The server has no auxiliary information about the private client data distribution. For completeness, our evaluation includes attacks that assume access to a public dataset aligned with the private data distribution (Li et al., 2022; Fang et al., 2023; Jeon et al., 2021).

2.2 DIFFUSION-BASED ATTACK FORMULATION

Diffusion Models. Diffusion models learn to reverse a corruption process that gradually transforms data into Gaussian noise (Ho et al., 2020; Song et al., 2020). Let $x_0 \sim p_{\text{data}}$ denote a clean sample. We use a pretrained diffusion model as the generative prior and DDIM for reverse sampling (Song et al., 2020). At timestep $t \in \{1, \ldots, T\}$, where T denotes the terminal diffusion timestep, the noisy state is $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ is determined by the noise schedule $\{\alpha_s \in (0,1)\}_{s=1}^{T}$. The denoising network $\epsilon_\phi(x_t, t)$ predicts the added noise by minimizing $\mathcal{L}_{\text{denoise}} = \mathbb{E}_{t,x_0,\epsilon}\big[\|\epsilon - \epsilon_\phi(x_t, t)\|_2^2\big]$, where ϕ denotes the network parameters. At each reverse step, the denoising network estimates the clean image as $\hat{x}_0(x_t) = \big(x_t - \sqrt{1-\bar{\alpha}_t}\, \epsilon_\phi(x_t, t)\big)/\sqrt{\bar{\alpha}_t}$. To steer the reverse process toward a desired condition, loss-guided diffusion introduces a guidance signal y and models the posterior distribution $p(x_t \mid y)$ (Guo et al., 2026; Yang et al., 2024). By Bayes' rule, the conditional score decomposes as

$$\nabla_{x_t} \log p(x_t \mid y) = \nabla_{x_t} \log p(x_t) + \nabla_{x_t} \log p(y \mid x_t). \tag{1}$$

The standard DDIM update represents the unconditional prior term $\nabla_{x_t} \log p(x_t)$, while the likelihood term $\nabla_{x_t} \log p(y \mid x_t)$ guides the sample toward consistency with y. We express the latter

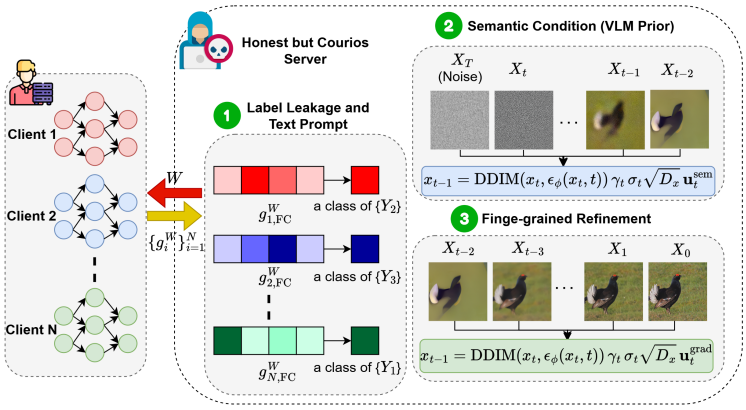


Figure 2: Overview of SEM2PIX. **(1) Label leakage:** the magnitude-aware score infers the private label set from the shared client gradient of the final FC layer, and each inferred label is used to construct a target text prompt. **(2) Vision–language semantic guidance:** during the early reverse steps ($t = T, \ldots, T/2$), a pretrained vision–language model aligns the diffusion trajectory with the semantic information encoded in the target prompt, establishing class-compatible global structure. **(3) Fine-grained refinement:** during the remaining steps ($t = T/2 - 1, \ldots, 1$), stage-wise gradient guidance dynamically matches the layer-wise gradients induced by the predicted clean image with the corresponding client gradients to recover fine-grained details.

using the guidance loss $\mathcal{L}_y(x_t) := -\log p(y \mid x_t)$, such that $\nabla_{x_t} \log p(y \mid x_t) = -\nabla_{x_t} \mathcal{L}_y(x_t)$. The resulting loss-guided DDIM update is

$$x_{t-1} = \mathrm{DDIM}(x_t, \epsilon_\phi(x_t, t)) - \gamma_t \nabla_{x_t} \mathcal{L}_y(x_t), \tag{2}$$

where $\gamma_t \geq 0$ controls the guidance strength. In diffusion-based gradient inversion, y corresponds to the shared client gradient $\mathbf{g}_{\mathrm{client}}$, and $\mathcal{L}_y$ is a differentiable gradient-matching objective. Minimizing $\mathcal{L}_y$ provides a guidance direction that steers reverse diffusion toward images consistent with the shared client gradient.

Gradient-Inversion Objective. Let f_θ denote the global classifier with parameters θ, and let $\mathcal{L}_{\mathrm{cls}}$ denote its training loss. For an image–label pair (x, k), the induced model gradient is $\mathbf{g}(x, k) := \nabla_\theta \mathcal{L}_{\mathrm{cls}}(f_\theta(x), k)$. The server observes the shared client gradient $\mathbf{g}_{\mathrm{client}}$. Let $\hat{k}$ denote a target class label inferred from $\mathbf{g}_{\mathrm{client}}$, as to be detailed in Section 3.2. Gradient inversion reconstructs an image by minimizing the discrepancy between its induced gradient and $\mathbf{g}_{\mathrm{client}}$ (Zhu et al., 2019; Wu et al., 2023). Following GGSS-R (Meng et al., 2026), we evaluate the discrepancy on the predicted clean image and formulate the reverse step as:

$$x_{t-1}^\star = \underset{x_{t-1} \in \mathcal{C}_{1-\delta}(x_t)}{\arg\min} \left[\nabla_{x_t} \left\| \mathbf{g}\left(\hat{x}_0(x_t), \hat{k}\right) - \mathbf{g}_{\mathrm{client}} \right\|_2 \right]^\top (x_{t-1} - x_t), \tag{3}$$

where $\delta \in (0, 1)$ and $\mathcal{C}_{1-\delta}(x_t)$ denote the $(1-\delta)$-confidence region of the reverse-transition distribution. This objective selects a high-probability reverse step aligned with the steepest decrease in gradient discrepancy. However, gradient consistency alone does not explicitly incorporate semantic information about the target class. This limitation motivates the vision–language semantic guidance introduced in Section 3.3.

3 PROPOSED METHOD: SEM2PIX

3.1 OVERVIEW

To effectively steer the reverse process of diffusion-based gradient inversion, we propose a two-stage guidance mechanism comprising vision–language semantic guidance and fine-grained refinement. The first stage constructs target text prompts from the labels inferred by the magnitude-aware score (Sec. 3.2). A pretrained vision–language model uses the semantic information encoded in these

prompts to guide the reverse trajectory. This stage recovers class-compatible global structure from $t = T$ to $t = T/2$, where T denotes the terminal diffusion timestep (Sec. 3.3). The second stage replaces semantic guidance with stage-wise gradient guidance for reverse steps $t = T/2 - 1, \ldots, 1$ to produce the reconstructed image x_0 (Sec. 3.4). The stage-wise guidance dynamically matches the layer-wise gradients induced by the predicted clean image to the corresponding client gradients and gradually incorporates earlier feature-extraction layers to recover fine-grained details. Figure 2 summarizes this semantic-to-pixel reconstruction process, and Algorithm 1 in Sec. A.2 of Appendix outlines SEM2PIX.

3.2 LABEL LEAKAGE VIA MAGNITUDE-AWARE INFERENCE

As noted in the first observation in Section 1.2, label leakage plays a crucial role in semantic recovery. For a batch of B samples drawn from K classes, let $\mathcal{K}_B^{\text{true}} = \{k_i^{\text{true}}\}_{i=1}^B$ denote the true private label set, where $k_i^{\text{true}} \in \{1, \ldots, K\}$ is the true class index of sample i. Consider the final fully connected (FC) layer with weight matrix $W_{\text{fc}} \in \mathbb{R}^{K \times d}$ and bias $\mathbf{b} \in \mathbb{R}^K$, where d is the feature dimension. For sample i, let $\mathbf{h}_i \in \mathbb{R}^d_{\geq 0}$ denote its feature vector and let $\mathbf{p}_i = \text{softmax}(W_{\text{fc}}\mathbf{h}_i + \mathbf{b})$ denote its class-probability vector. Under the softmax cross-entropy loss, row k of the unperturbed FC weight gradient $\mathbf{g}_{\text{client}}^{\text{FC}} \in \mathbb{R}^{K \times d}$ is

$$\mathbf{g}_{\text{client},k,:}^{\text{FC}} = \frac{1}{B} \sum_{i=1}^{B} \left(p_{i,k} - \mathbb{I}(k_i^{\text{true}} = k)\right) \mathbf{h}_i^\top, \tag{4}$$

where $\mathbf{g}_{\text{client}}^{\text{FC}}$ denotes the FC component of the shared client gradient $\mathbf{g}_{\text{client}}$, $p_{i,k}$ is the probability assigned to class $k \in \{1, \ldots, K\}$, and $\mathbb{I}(\cdot)$ is the indicator function. Prior work infers private labels by selecting classes whose FC-gradient rows contain only negative values (Zhao et al., 2020; Ye et al., 2024). However, defense perturbations can flip the signs of entries within an FC-gradient row (Hu et al., 2026). We therefore measure the magnitude of negative gradient values associated with each class. Let $\mathbf{g}_{\text{client}}^{\text{FC(perturb)}} = \mathbf{g}_{\text{client}}^{\text{FC}} + \mathbf{n}^{W_{\text{fc}}}$ denote the perturbed FC component, where $\mathbf{n}^{W_{\text{fc}}} \in \mathbb{R}^{K \times d}$ represents the defense-induced perturbation. We define the magnitude-aware score as:

$$S_k^{\text{mag}} \coloneqq \left\| \left[-\mathbf{g}_{\text{client},k,:}^{\text{FC(perturb)}} \right]_+ \right\|_1, \qquad k \in \{1, \ldots, K\}, \tag{5}$$

where $[\cdot]_+$ is the elementwise positive-part operator, and a larger score indicates stronger evidence that the corresponding class is present in the private batch. Given the known client batch size B (Fan et al., 2025b), we define the inferred private label set using the B highest-scoring classes:

$$\widehat{\mathcal{K}}_B \coloneqq \{\hat{k}_1, \ldots, \hat{k}_B\}, \qquad S_{\hat{k}_1}^{\text{mag}} \geq S_{\hat{k}_2}^{\text{mag}} \geq \cdots \geq S_{\hat{k}_B}^{\text{mag}}, \tag{6}$$

where $\hat{k}_i$ denotes the inferred class index at rank $i \in \{1, \ldots, B\}$. For $B = 1$, this selects the class with the largest magnitude-aware score.

3.3 VISION–LANGUAGE SEMANTIC GUIDANCE

After inferring the private label set in Section 3.2, we construct text prompts to guide the reverse process of a training-free diffusion model, following 2nd observation in Section 1.2. For each reconstruction trajectory, we select a target label $\hat{k} \in \widehat{\mathcal{K}}_B$. We remark that the reverse process of a training-free diffusion model cannot be guided directly by a discrete label because updating the reverse diffusion trajectory requires an objective that is differentiable with respect to the noisy state x_t(Guo et al., 2026; 2024; Yang et al., 2024).

In view of this limitation, we construct a natural-language prompt τ from the inferred label $\hat{k}$, such as "an image of a [class name]." A pretrained vision–language model then maps text and images into a shared embedding space, thereby providing a differentiable objective that aligns the predicted clean image with the semantic information encoded in the target prompt. This objective provides training-free semantic guidance to a training-free diffusion model. We instantiate the vision–language model with CLIP (Radford et al., 2021; Zhang et al., 2025b; Kim et al., 2026), and define the semantic loss and its normalized gradient using the CLIP image and text encoders, E_I and E_T, respectively, as

$$\mathcal{L}_{\text{sem}}(x_t, \tau) = 1 - \frac{E_I(\hat{x}_0(x_t))^\top E_T(\tau)}{\|E_I(\hat{x}_0(x_t))\|_2 \, \|E_T(\tau)\|_2}, \qquad \mathbf{u}_t^{\text{sem}} = \frac{\nabla_{x_t} \mathcal{L}_{\text{sem}}(x_t, \tau)}{\|\nabla_{x_t} \mathcal{L}_{\text{sem}}(x_t, \tau)\|_2}. \tag{7}$$

Following the objective of DDIM update in Eq. (2), the direction $-\mathbf{u}_t^{\text{sem}}$ decreases the semantic loss and aligns the predicted clean image with the semantic information encoded in the target prompt. We then apply normalized noise-dependent guidance (Yang et al., 2024) to correct the reverse transition as:

$$x_{t-1} = \text{DDIM}(x_t, \epsilon_\phi(x_t, t)) - \gamma_t \, \sigma_t \sqrt{D_x} \, \mathbf{u}_t^{\text{sem}}, \tag{8}$$

where $\sigma_t = \sqrt{1 - \bar{\alpha}_t}$ denotes the diffusion noise level at timestep t and D_x denotes the number of image elements. The factor $\sigma_t \sqrt{D_x}$ scales the normalized guidance correction according to the diffusion noise level and image dimensionality. This vision–language semantic correction retains the pretrained DDIM prior and steers reverse sampling using the semantic information encoded in the target prompt.

3.4 Stage-Wise Gradient Guidance

Training-free diffusion guidance steers reverse diffusion using gradients of differentiable objectives (Guo et al., 2026; 2024; Yang et al., 2024). For high-resolution image reconstruction in FL, gradient matching can be computationally expensive and may struggle to recover fine-grained details (Ye et al., 2024). We introduce stage-wise gradient guidance that progressively incorporates layer-wise gradient information to refine the reconstructed images.

For a model with L layers, let θ_ℓ and $\mathbf{g}_{\text{client}}^{(\ell)}$ denote the parameters and shared client gradient of layer $\ell \in \{1, \ldots, L\}$, respectively. At timestep t, we define the active layer set as $\mathcal{A}_t = \{L - s_t + 1, \ldots, L\}$, where s_t denotes the number of active layers and progressively increases from 1 to L, expanding the active layer set from the final fully connected (FC) layer toward earlier feature-extraction layers. Following semantic guidance (Section 3.3), the noisy batch $\mathbf{X}_t = \{x_{t,i}\}_{i=1}^B$ yields predicted clean images $\hat{x}_{0,i}(x_{t,i})$, paired with fixed inferred labels $\hat{k}_i$. We define their batch-averaged gradient at layer ℓ as $\widehat{\mathbf{g}}_t^{(\ell)} \coloneqq \frac{1}{B} \sum_{i=1}^B \nabla_{\theta_\ell} \mathcal{L}_{\text{cls}}\Big(f_\theta(\hat{x}_{0,i}(x_{t,i})), \hat{k}_i\Big)$. With the labels and active layers fixed at timestep t, the gradient-matching loss is

$$\mathcal{L}_{\text{grad}}(\mathbf{X}_t) = \sum_{\ell \in \mathcal{A}_t} \omega_\ell \left\| \widehat{\mathbf{g}}_t^{(\ell)} - \mathbf{g}_{\text{client}}^{(\ell)} \right\|_2^2, \tag{9}$$

where $\omega_\ell \geq 0$ weights the contribution of layer ℓ. This loss couples all B reconstruction trajectories through the batch-averaged gradient. The normalized guidance direction for trajectory i is

$$\mathbf{u}_{t,i}^{\text{grad}} = \frac{\nabla_{x_{t,i}} \mathcal{L}_{\text{grad}}(\mathbf{X}_t)}{\left\| \nabla_{x_{t,i}} \mathcal{L}_{\text{grad}}(\mathbf{X}_t) \right\|_2}, \qquad i \in \{1, \ldots, B\}. \tag{10}$$

At each refinement step, we compute all guidance directions from the same batch $\mathbf{X}_t$ before updating any trajectory. The direction $-\mathbf{u}_{t,i}^{\text{grad}}$ then steers trajectory i to reduce the gradient-matching loss, jointly refining fine-grained details across the reconstructed images.

3.5 Semantic-to-Pixel Guidance

Sem2Pix activates the two guidance mechanisms at different steps of reverse diffusion. For $t = T, \ldots, T/2$, vision–language semantic guidance (Sec. 3.3) steers the trajectory toward the semantic information encoded in the target prompt. For the remaining steps, $t = T/2 - 1, \ldots, 1$, stage-wise gradient guidance (Section 3.4) replaces semantic guidance to refine fine-grained details. For each trajectory $i \in \{1, \ldots, B\}$, let τ_i denote the prompt constructed from $\hat{k}_i$, and let $\mathbf{u}_{t,i}^{\text{sem}}$ denote the semantic direction in Eq. equation 7 evaluated at $(x_{t,i}, \tau_i)$. The active guidance direction for trajectory i is then defined as:

$$\mathbf{u}_{t,i}^{\text{S2P}} = \begin{cases} \mathbf{u}_{t,i}^{\text{sem}}, & T/2 \leq t \leq T, \\ \mathbf{u}_{t,i}^{\text{grad}}, & 1 \leq t < T/2. \end{cases} \tag{11}$$

During fine-grained refinement, the active layer set expands from the final FC layer to earlier feature-extraction layers. We apply the active guidance direction to each trajectory as:

$$x_{t-1,i} = \text{DDIM}(x_{t,i}, \epsilon_\phi(x_{t,i}, t)) - \gamma_t \, \sigma_t \sqrt{D_x} \, \mathbf{u}_{t,i}^{\text{S2P}}. \tag{12}$$

At each reverse step, we adjust the DDIM output using the active guidance direction. Semantic guidance first helps recover the global shape and appearance described by the target prompt. Stage-wise gradient guidance then uses the shared client gradients to refine image details. This process produces the reconstructed batch $\{x_{0,i}\}_{i=1}^{B}$.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Evaluation protocol. We used a default configuration with ResNet-18 He et al. (2016), a client batch size of 1, and one local training step. We conducted experiments on three natural-image datasets, *i.e.*, CelebA (Karras et al., 2018), LSUN (Yu et al., 2015), and ImageNet (Deng et al., 2009). We also evaluated SEM2PIX on the privacy-sensitive healthcare imaging dataset Camelyon17 (Bandi et al., 2018). We used the pretrained diffusion model from Chung et al. (Chung et al., 2023), which was trained on FFHQ at a resolution of 256×256 pixels.

Hyperparameters. We set the terminal diffusion timestep to $T = 1000$ and used the semantic-to-pixel reverse update in Eq. (12) with a fixed guidance strength of $\gamma_t = 0.30$ at all reverse steps.

Evaluation metrics. We evaluated reconstruction quality using LPIPS, SSIM, and PSNR. We further measured semantic fidelity using CLIP similarity (Radford et al., 2021) and fine-grained structural quality using GMSD (Gradient Magnitude Similarity Deviation) (Xue et al., 2013).

Baselines. We compared SEM2PIX with state-of-the-art GIAs from three categories: (i) Pixel-based baselines are ARES (Gong et al., 2026), FedLeak (Fan et al., 2025b), and HFGradInv (Ye et al., 2024); (ii) GAN-based baselines are GGL (Li et al., 2022), GIFD (Fang et al., 2023), and GIAS (Jeon et al., 2021); and (iii) Diffusion-based baselines are GGDM (Gu et al., 2024) and GGSS-R (Meng et al., 2026). We omit a few gradient inversion attacks (Wan et al., 2026; Carletti et al., 2026) from our evaluation due to unreleased codebases, as well as GLAD (Xiang et al., 2026) due to the unavailability of its pre-trained generator checkpoint. Following (Carletti et al., 2025), our ablation studies progressively examined more challenging configurations.

4.2 EXPERIMENTAL RESULTS

4.2.1 RECONSTRUCTION FROM UNPERTURBED GRADIENTS

Table 1 compares SEM2PIX with state-of-the-art GIAs using unperturbed gradients across the evaluated datasets. We highlight three findings: (i) GAN-based attacks exhibit suboptimal performance, which we attribute in part to the distribution mismatch between private images and auxiliary data used to train the GAN models. (ii) Recent pixel-based attacks, such as FedLeak (Fan et al., 2025b) and HFGradInv (Ye et al., 2024), and diffusion-based attacks show promising performance. (iii) SEM2PIX outperforms GGDM (Gu et al., 2024) and GGSS-R (Meng et al., 2026) across the reported metrics, except for PSNR on CelebA and LSUN, where GGSS-R performs better. SEM2PIX uses vision–language guidance to align reconstruction with the semantic information encoded in the target text prompt and stage-wise gradient guidance to progressively match layer-wise gradients and recover fine-grained details. The qualitative results in Fig. 3a show that SEM2PIX preserves both class-compatible global structure and fine-grained visual details. A detailed attack cost analysis of SEM2PIX can be found in Sec. A.3 of the Appendix.

4.2.2 RECONSTRUCTION AGAINST DEFENSE MECHANISMS

We evaluate the robustness of SEM2PIX on ImageNet under three gradient defenses: Gaussian noise ($\sigma_{\text{noise}} = 0.1$), gradient clipping ($C = 0.5C_{\text{med}}$), where C_{med} denotes the median per-example gradient norm, and Soteria with a 60% feature-pruning rate (Sun et al., 2021). More discussion on the utility-privacy trade-off under varying defense strengths can be found in Sec. A.4 of the Appendix. Table 2 summarizes the reconstruction performance.

We highlight two findings: (i) Pixel-based attacks are sensitive to perturbed gradients. (ii) SEM2PIX outperforms the state-of-the-art baselines under all evaluated defenses. Gradient perturbations can

Table 1: Reconstruction performance from unperturbed gradients.

Dataset	Category	Method	LPIPS ↓	SSIM ↑	PSNR ↑	GMSD ↓	CLIP similarity ↑
CelebA	Pixel-based	HFGradInv (Ye et al., 2024)	0.2576	0.4574	15.4752	0.2165	0.7283
		FedLeak (Fan et al., 2025b)	0.4206	0.6034	19.8542	0.1985	0.7477
		ARES (Gong et al., 2026)	0.4821	0.5412	17.6521	0.2245	0.7180
	GAN-based	GIAS (Jeon et al., 2021)	1.0021	0.3395	10.1563	0.2931	0.2272
		GGL (Li et al., 2022)	0.8054	0.2966	10.9504	0.2609	0.4457
		GIFD (Fang et al., 2023)	1.0057	0.3411	9.8926	0.2807	0.2774
		GGDM (Gu et al., 2024)	0.2303	0.5524	19.418	0.1918	0.7832
	Diffusion-based	GGSS-R (Meng et al., 2026)	0.0272	0.9269	**38.513**	0.0124	0.9634
		SEM2PIX	**0.0061**	**0.9366**	37.904	**0.0056**	**0.9835**
ImageNet	Pixel-based	HFGradInv (Ye et al., 2024)	0.3621	0.3253	16.4015	0.2130	0.7648
		FedLeak (Fan et al., 2025b)	0.5992	0.4842	19.7427	0.2048	0.7219
		ARES (Gong et al., 2026)	0.6215	0.4120	16.8920	0.2312	0.6985
	GAN-based	GIAS (Jeon et al., 2021)	0.9449	0.3523	12.8365	0.2615	0.2680
		GGL (Li et al., 2022)	0.7769	0.3549	13.8770	0.2619	0.7357
		GIFD (Fang et al., 2023)	0.8419	0.3558	12.9575	0.2651	0.2950
		GGDM (Gu et al., 2024)	0.6991	0.3308	17.1600	0.3016	0.2286
	Diffusion-based	GGSS-R (Meng et al., 2026)	0.1787	0.5617	22.3760	0.0195	0.9561
		SEM2PIX	**0.0776**	**0.7451**	**25.8470**	**0.0086**	**0.9864**
LSUN	Pixel-based	HFGradInv (Ye et al., 2024)	0.3757	0.3554	13.9454	0.2451	0.6954
		FedLeak (Fan et al., 2025b)	0.5618	0.6194	18.4373	0.1906	0.6380
		ARES (Gong et al., 2026)	0.5894	0.4865	15.6240	0.2384	0.6421
	GAN-based	GIAS (Jeon et al., 2021)	0.9080	0.4097	10.2932	0.2695	0.2462
		GGL (Li et al., 2022)	0.8261	0.3844	10.6909	0.2609	0.4457
		GIFD (Fang et al., 2023)	0.9563	0.3761	9.02370	0.3015	0.2603
		GGDM (Gu et al., 2024)	0.5051	0.4907	20.1170	0.2147	0.2242
	Diffusion-based	GGSS-R (Meng et al., 2026)	0.0737	0.9148	**35.2980**	0.0285	0.9671
		SEM2PIX	**0.0197**	**0.9322**	33.6680	**0.0043**	**0.9888**

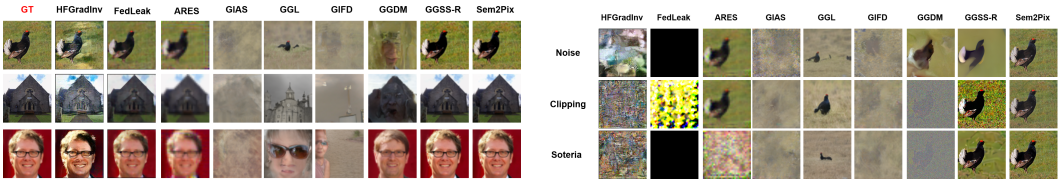
Table 2: Reconstruction performance on ImageNet under three gradient defenses.

Defense	Category	Method	LPIPS ↓	SSIM ↑	PSNR ↑	GMSD ↓	CLIP similarity ↑
Noise	Pixel-based	HFGradInv (Ye et al., 2024)	0.8348	0.1756	9.8160	0.2611	0.2403
		FedLeak (Fan et al., 2025b)	1.0939	0.0014	7.3521	0.2898	0.2253
		ARES (Gong et al., 2026)	0.7124	0.4215	15.4210	0.2214	0.6845
	GAN-based	GIAS (Jeon et al., 2021)	0.9635	0.2953	12.9631	0.2258	0.2913
		GGL (Li et al., 2022)	0.8847	0.2904	13.0589	0.2551	0.7081
		GIFD (Fang et al., 2023)	0.9444	0.2895	12.8404	0.2438	0.2997
		GGDM (Gu et al., 2024)	0.7680	0.2912	13.1123	0.2431	0.6394
	Diffusion-based	GGSS-R (Meng et al., 2026)	0.6808	0.4104	18.6181	0.1542	0.9041
		SEM2PIX	**0.2650**	**0.6091**	**23.0243**	**0.0882**	**0.9793**
Gradient clipping	Pixel-based	HFGradInv (Ye et al., 2024)	0.7686	0.0831	12.7126	0.2698	0.2117
		FedLeak (Fan et al., 2025b)	0.8653	0.1403	7.1293	0.2908	0.2224
		ARES (Gong et al., 2026)	0.7052	0.4310	15.8924	0.2185	0.6912
	GAN-based	GIAS (Jeon et al., 2021)	0.8410	0.3663	13.5966	0.2592	0.2665
		GGL (Li et al., 2022)	0.7651	0.3435	13.0741	0.2469	0.7946
		GIFD (Fang et al., 2023)	0.8479	0.3612	13.6111	0.2555	0.3744
		GGDM (Gu et al., 2024)	0.4821	0.3415	17.4912	0.1821	0.5487
	Diffusion-based	GGSS-R (Meng et al., 2026)	0.1673	0.5648	22.6929	0.1076	0.9597
		SEM2PIX	**0.0600**	**0.8140**	**27.4739**	**0.0485**	**0.9809**
Soteria	Pixel-based	HFGradInv (Ye et al., 2024)	0.8858	0.0640	12.3281	0.2711	0.1853
		FedLeak (Fan et al., 2025b)	1.0939	0.0014	7.3521	0.2898	0.2253
		ARES (Gong et al., 2026)	0.9842	0.1124	9.4512	0.2810	0.2541
	GAN-based	GIAS (Jeon et al., 2021)	0.8435	0.3694	13.8393	0.2576	0.2749
		GGL (Li et al., 2022)	0.7984	0.3503	13.4761	0.2529	0.6442
		GIFD (Fang et al., 2023)	0.8123	0.3576	13.6097	0.2594	0.2979
		GGDM (Gu et al., 2024)	0.6125	0.3462	16.8931	0.2158	0.5498
	Diffusion-based	GGSS-R (Meng et al., 2026)	0.5092	0.5292	23.1669	0.1725	0.7653
		SEM2PIX	**0.1084**	**0.6638**	**25.4587**	**0.0809**	**0.9719**

weaken the reconstruction signal across layers. SEM2PIX leverages the generative prior of the diffusion model and applies stage-wise gradient guidance to match the layer-wise gradients induced by the reconstructed batch with the corresponding perturbed client gradients. Matching begins with the final FC layer and progressively incorporates earlier feature-extraction layers. Fig. 3b shows that SEM2PIX preserves object structure and local appearance across all three defenses, supporting its robustness to perturbed client gradients.

4.2.3 RECONSTRUCTION IN A PRIVACY-SENSITIVE HEALTHCARE DOMAIN

We further evaluated SEM2PIX on Camelyon17, a privacy-sensitive histopathology dataset containing fine-grained tissue structures and local morphological patterns that may reveal patient-specific information. Table 3 shows that SEM2PIX outperforms the evaluated baselines in terms of LPIPS,



(a) Reconstruction from unperturbed client gradients.

(b) Reconstruction from perturbed gradients.

Figure 3: Comparison of qualitative reconstruction results.

Table 3: Reconstruction performance on Camelyon17.

Attack type	Method	LPIPS ↓	SSIM ↑	PSNR ↑	GMSD ↓	CLIP similarity ↑
Pixel-based	HFGradInv (Ye et al., 2024)	0.6227	0.2277	10.2884	0.2490	0.7672
	FedLeak (Fan et al., 2025b)	0.3611	0.7432	16.9975	0.1818	0.8152
	ARES (Gong et al., 2026)	0.4625	0.5842	14.8520	0.2014	0.7891
GAN-based	GIAS (Jeon et al., 2021)	0.9117	0.3907	8.3269	0.3344	0.5238
	GGL (Li et al., 2022)	0.8823	0.3028	7.3784	0.2804	0.4175
	GIFD (Fang et al., 2023)	0.9241	0.3866	8.3732	0.3319	0.5430
Diffusion-based	GGDM (Gu et al., 2024)	1.3899	0.0572	12.9389	0.2906	0.3146
	GGSS-R (Meng et al., 2026)	0.3713	0.3265	10.2890	0.2130	0.4402
	SEM2PIX	**0.1881**	**0.6708**	**16.1124**	**0.1289**	**0.8322**

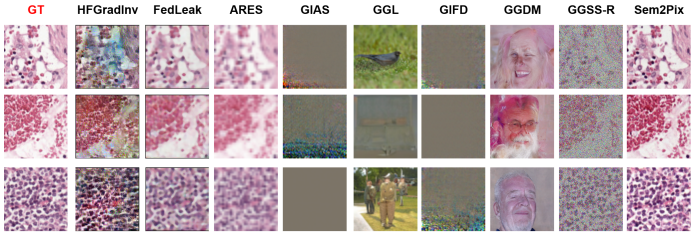


Figure 4: Qualitative reconstruction results on the privacy-sensitive Camelyon17 healthcare dataset.

GMSD, and CLIP, while FedLeak achieves higher SSIM and PSNR. Figure 4 further shows that SEM2PIX preserves tissue structure and local texture in healthcare images, supporting the effectiveness of the semantic-to-pixel strategy in this domain.

5 MORE EXPERIMENTAL EVALUATION IN THE APPENDIX

To support our study, we provide additional experimental evaluations in the appendix. For a more practical FL setting Carletti et al. (2025), we evaluated reconstruction performance under larger client batch sizes, multiple local update steps, different communication rounds, and deeper model architectures in Sec. A.5 of the Appendix. We conducted ablation studies on the guidance mechanisms, label inference, and guidance strength in Sec. A.6 of the Appendix.

6 CONCLUSION

In this paper, we hypothesize that effective gradient inversion attacks should recover both semantic and fine-grained details of private images. Based on this insight, we present SEM2PIX, a training-free diffusion-based GIA with a two-stage guidance mechanism. First, vision–language semantic guidance steers reverse diffusion toward a global shape and color distribution consistent with the target text prompt. Second, stage-wise gradient guidance replaces semantic guidance and progressively matches layer-wise gradients to refine fine-grained details. With these two steps, our experiments confirm strong reconstruction performance from unperturbed gradients and in more practical FL settings.

AI USE STATEMENT

We used generative AI tools, including large language models (LLMs), to assist with language editing and improve the clarity of the manuscript. The authors reviewed and revised all AI-assisted text to ensure its accuracy, clarity, and consistency with the research presented. All LLM-generated code was verified and tested for correctness by the authors. The authors take full responsibility for the final manuscript and accompanying code.

ETHICS STATEMENT

This work investigates privacy leakage in federated learning through gradient inversion attacks, with the aim of assessing privacy risks and supporting the development of stronger privacy-preserving systems. Our experiments used established benchmark datasets and did not involve the collection of new data from human participants. We acknowledge the dual-use implications of gradient inversion techniques, which could be misused to reconstruct sensitive client data without authorization. We present this method for research and privacy evaluation to advance understanding of these vulnerabilities and inform the development of more effective defenses. We do not endorse its use against systems or data without appropriate authorization.

REPRODUCIBILITY STATEMENT

Our implementation was built on the GGSS-R codebase (Meng et al., 2026), available at `https://github.com/mjyyhhxx/GGSS-R`. We implemented the state-of-the-art attack baselines using their officially released code. For vision–language semantic guidance, we used a CLIP-based objective (Radford et al., 2021) with the implementation available at `https://github.com/ExcelsiorCJH/CLIP`. For the diffusion prior, we used the pretrained model trained on FFHQ at a resolution of 256×256 pixels, provided by (Chung et al., 2023). The corresponding code and pretrained weights are available at `https://github.com/DPS2022/diffusion-posterior-sampling`.

REFERENCES

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

Vincenzo Carletti, Pasquale Foggia, Carlo Mazzocca, Giuseppe Parrella, and Mario Vento. {SoK}: Gradient inversion attacks in federated learning. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 6439–6459, 2025.

Vincenzo Carletti, Pasquale Foggia, Carlo Mazzocca, Giuseppe Parrella, and Mario Vent. Guide: Enhancing gradient inversion attacks in federated learning with denoising models. *IEEE Internet of Things Journal*, 2026.

Sherry X. Chen, Misha Sra, and Pradeep Sen. Instruct-clip: Improving instruction-guided image editing with automated data refinement using contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28513–28522, June 2025.

Hyungjin Chung, Jeongsol Kim, Michael T. Mccann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Ivan Costa, Pedro Correia, Ivone Amorim, Eva Maia, and Isabel Praça. Towards privacy-preserving federated learning using hybrid homomorphic encryption. In *International Conference on Applied Cryptography and Network Security*, pp. 105–128. Springer, 2026.

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, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Mingyuan Fan, Cen Chen, Chengyu Wang, Xiaodan Li, and Wenmeng Zhou. Refiner: Data refining against gradient leakage attacks in federated learning. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3005–3024, 2025a.

Mingyuan Fan, Fuyi Wang, Cen Chen, and Jianying Zhou. Boosting gradient leakage attacks: Data reconstruction in realistic fl settings. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 2985–3004, 2025b.

Hao Fang, Bin Chen, Xuan Wang, Zhi Wang, and Shu-Tao Xia. Gifd: A generative gradient inversion method with feature domain optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4967–4976, 2023.

Zirui Gong, Leo Yu Zhang, Yanjun Zhang, Viet Vo, Tianqing Zhu, Shirui Pan, and Cong Wang. Ares: Scalable and practical gradient inversion attack in federated learning through activation recovery. In *2026 IEEE Symposium on Security and Privacy (SP)*, pp. 2327–2345. IEEE, 2026.

Hongyan Gu, Xinyi Zhang, Jiang Li, Hui Wei, Baiqi Li, and Xinli Huang. Federated learning vulnerabilities: Privacy attacks with denoising diffusion probabilistic models. In *Proceedings of the ACM Web Conference 2024*, pp. 1149–1157, 2024.

Yingqing Guo, Hui Yuan, Yukang Yang, Minshuo Chen, and Mengdi Wang. Gradient guidance for diffusion models: An optimization perspective. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yingqing Guo, Yukang Yang, Hui Yuan, and Mengdi Wang. Training-free guidance beyond differentiability: Scalable path steering with tree search in diffusion and flow models. *Advances in Neural Information Processing Systems*, 38:73343–73384, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

Bin Hu, Jingling Yuan, Jiawei Jiang, and Chuang Hu. Venom: Liquid diffusion-guided gradient inversion for breaking differential privacy in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21814–21822, 2026.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2261–2269. Ieee, 2017.

Jinwoo Jeon, Kangwook Lee, Sewoong Oh, Jungseul Ok, et al. Gradient inversion with generative image prior. *Advances in neural information processing systems*, 34:29898–29908, 2021.

Peter Kairouz and H Brendan McMahan. Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1-2):1–210, 2021.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=Hk99zCeAb.

Dain Kim, Jinseo Kim, and Sungyong Baik. Training-free debiasing of diffusion models via clip-guided denoising optimization. *arXiv preprint arXiv:2607.00817*, 2026.

Zhuohang Li, Jiaxin Zhang, Luyang Liu, and Jian Liu. Auditing privacy defenses in federated learning via generative gradient leakage. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10132–10142, 2022.

Xuan Liu, Siqi Cai, Qihua Zhou, Song Guo, Ruibin Li, and Kaiwei Lin. Mjölnir: Breaking the shield of perturbation-protected gradients via adaptive diffusion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 26308–26316, 2025.

Yunlv Lv, Rui Zhang, Zhiyuan Zhang, Ziyi Wan, Lanxue Zhang, Minhui Xue, Jiangtao Li, and Yanan Cao. SMASH: Scalable maliciously secure hybrid multi-party computation framework for privacy-preserving large language models. In *Proceedings of the 35th USENIX Security Symposium (USENIX Security 26)*. USENIX Association, 2026. URL `https://usenix.org`.

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.

Jiayang Meng, Tao Huang, Hong Chen, and Cuiping Li. Is diffusion model safe? severe data leakage via gradient-guided diffusion model. *arXiv preprint arXiv:2406.09484*, 2024.

Jiayang Meng, Tao Huang, Hong Chen, Chen Hou, and Guolong Zheng. Enhanced privacy leakage from noise-perturbed gradients via gradient-guided conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7981–7989, 2026.

Pinelopi Papalampidi, Olivia Wiles, Ira Ktena, Aleksandar Shtedritski, Emanuele Bugliarello, Ivana Kajić, Isabela Albuquerque, and Aida Nematzadeh. Dynamic classifier-free diffusion guidance via online feedback. In *International Conference on Learning Representations*, volume 2026, pp. 127362–127378, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Farchan Hakim Raswa, Chun-Shien Lu, and Jia-Ching Wang. Histofs: Non-iid histopathologic whole slide image classification via federated style transfer with roi-preserving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30251–30260, 2025.

Junjie Shan, Ziqi Zhao, Jialin Lu, Rui Zhang, Siu Ming Yiu, and Ka-Ho Chow. Geminio: Language-guided gradient inversion attacks in federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2718–2727, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Jingwei Sun, Ang Li, Binghui Wang, Huanrui Yang, Hai Li, and Yiran Chen. Soteria: Provable defense against privacy leakage in federated learning from representation perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9311–9319, 2021.

Cheng-Chang Tsai, Kai-Wen Cheng, and Chun-Shien Lu. Fedsda: Federated stain distribution alignment for non-iid histopathological image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9566–9575, 2026.

Guangnian Wan, Gongfan Fang, Xinyin Ma, and Xinchao Wang. Gradient inversion attacks beyond sgd. In *Forty-third International Conference on Machine Learning*, 2026.

Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE transactions on information forensics and security*, 15:3454–3469, 2020.

Ruihan Wu, Xiangyu Chen, Chuan Guo, and Kilian Q Weinberger. Learning to invert: Simple adaptive attacks for gradient inversion in federated learning. In *Uncertainty in Artificial Intelligence*, pp. 2293–2303. PMLR, 2023.

Kunlan Xiang, Haomiao Yang, Meng Hao, Zikang Ding, Hongwei Li, Qingchuan Zhao, and Tianwei Zhang. An advanced gradient leakage attack against duplicate labels via model outputs reconstruction. *IEEE Transactions on Dependable and Secure Computing*, 2026.

Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE transactions on image processing*, 23(2): 684–695, 2013.

Jian Yang, Dacheng Yin, Xiaoxuan He, Yong Li, Fengyun Rao, Jing Lyu, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Wemmu: Enhanced bridging of vision-language models and diffusion models via noisy query tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17609–17618, 2026.

Lingxiao Yang, Shutong Ding, Yifan Cai, Jingyi Yu, Jingya Wang, and Ye Shi. Guidance with spherical gaussian constraint for conditional diffusion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 56071–56095. PMLR, 2024.

Zipeng Ye, Wenjian Luo, Qi Zhou, and Yubo Tang. High-fidelity gradient inversion in distributed learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19983–19991, 2024.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Kaiyuan Zhang, Siyuan Cheng, Guangyu Shen, Bruno Ribeiro, Shengwei An, Pin-Yu Chen, Xiangyu Zhang, and Ninghui Li. CENSOR: Defense against gradient inversion via orthogonal subspace bayesian sampling. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2025a.

Youyuan Zhang, Zehua Liu, Zenan Li, Zhaoyu Li, James J Clark, and Xujie Si. Decoupling training-free guided diffusion by admm. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23292–23302. IEEE, 2025b.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

A Appendix

A.1 Related Work

A.1.1 Gradient Inversion Attacks

Pixel-based attacks initialize dummy images and iteratively optimize them so that their induced gradients match the client gradient (Zhu et al., 2019; Wu et al., 2023; Ye et al., 2024; Wan et al., 2026; Fan et al., 2025b; Gong et al., 2026). HFGradInv (Ye et al., 2024) proposes a stepwise gradient-matching strategy for the high-fidelity reconstruction of high-resolution images. Fan et al. (Fan et al., 2025b) propose a two-stage recovery framework based on partial gradient matching and regularization. ARES (Gong et al., 2026) formulates gradient reconstruction as a noisy sparse recovery problem, enabling recovery from large batches. GAN-based attacks reduce the reconstruction search space by optimizing over the image manifold of a pretrained generator (Li et al., 2022; Fang et al., 2023; Jeon et al., 2021; Xiang et al., 2026). For example, GIFD (Fang et al., 2023) leverages intermediate feature representations of a pretrained GAN as optimization variables in a gradient-matching objective for image reconstruction. GLAD (Xiang et al., 2026) instead disentangles the feature maps of the target model using inferred labels and recovered model outputs, enabling the reconstruction of samples with duplicate labels. However, the effectiveness of these attacks depends on the alignment between the generator's pretraining distribution and the private client data distribution. More recently, diffusion models have emerged as alternative generative priors for gradient inversion. Early work by Meng et al. (Meng et al., 2024) adapts a pretrained diffusion prior to stabilize reconstruction. Subsequent work improves robustness to practical client updates and introduces measurement-guided reverse diffusion. Mjolnir (Liu et al., 2025) and Venom (Hu et al., 2026) address client updates perturbed by defense mechanisms such as noise. GGDM (Gu et al., 2024) and GGSS-R (Meng et al., 2026) use the client gradient as measurement guidance in the predicted-noise and predicted-clean-image spaces, respectively. GUIDE (Carletti et al., 2026) enhances the reconstruction quality of existing GIAs using denoising models. In a related direction, Geminio (Shan et al., 2025) introduces language-guided gradient inversion by modifying the global model and its loss surface so that the client update emphasizes training samples semantically aligned with an attacker-specified text query.

Remarks. Sem2Pix is training-free and employs a two-stage guidance strategy. Its semantic guidance stage constructs target text prompts from the inferred private label set and uses the semantic information encoded in these prompts to guide reverse diffusion. Its fine-grained refinement stage then applies stage-wise gradient guidance to progressively match layer-wise gradients. Unlike Geminio, Sem2Pix leaves the global model and the FL objective unchanged.

A.1.2 Language-Guided Diffusion Models

Diffusion models generate high-quality images by learning to reverse a gradual corruption process (Ho et al., 2020; Song et al., 2020). Guided diffusion methods steer this reverse process toward desired content using auxiliary guidance signals. Class-based methods incorporate label information through classifier-free guidance (Papalampidi et al., 2026; Ho & Salimans, 2021). Alternatively, vision–language models (VLMs) provide semantic guidance through natural-language supervision (Radford et al., 2021; Yang et al., 2026; Chen et al., 2025). Specifically, VLMs align textual and visual representations in a shared embedding space, enabling semantic information from language to guide the denoising process. However, training a target-specific guided diffusion model typically requires substantial labeled image–text data. More recently, training-free guidance methods have improved the applicability of pretrained diffusion models by steering them during inference without additional training. Many approaches formulate guidance through differentiable objectives and use their gradients to modify the reverse diffusion trajectory (Guo et al., 2026; 2024; Yang et al., 2024). Vision–language models further provide semantic guidance by defining text–image alignment objectives, thereby enabling semantically meaningful changes during the reverse diffusion process (Zhang et al., 2025b; Kim et al., 2026). This training-free formulation is particularly useful when insufficient data are available for training a target-specific guided model.

Remarks. Building on insights from training-free diffusion guidance, we study the effectiveness of vision–language semantic guidance and stage-wise gradient guidance in steering reconstructions

Algorithm 1 SEM2PIX: Semantic-to-Pixel Reconstruction

Require: Shared gradient $\mathbf{g}_{\text{client}}$; batch size B; classifier f_θ with K classes, L layers, and loss $\mathcal{L}_{\text{cls}}$
Require: Pretrained diffusion model ϵ_ϕ; vision–language encoders E_I, E_T; terminal timestep T; noise schedule $\{\bar{\alpha}_t\}$
Require: Guidance strengths $\{\gamma_t\}$; image dimensionality D_x; nonnegative layer weights $\{\omega_\ell\}$; active-layer counts $\{s_t\}_{t=1}^{T/2-1}$
Ensure: Reconstructed batch $\{x_{0,i}\}_{i=1}^{B}$

Infer $\widehat{\mathcal{K}}_B \leftarrow \{\hat{k}_1, \ldots, \hat{k}_B\}$ from the top-B scores S_k^{mag} according to Sec. 3.2
Construct prompts τ_i from $\hat{k}_i$ and initialize $x_{T,i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, T-1, \ldots, 1$ **do**
    Set $\mathbf{X}_t \leftarrow \{x_{t,i}\}_{i=1}^{B}$ and $\sigma_t \leftarrow \sqrt{1-\bar{\alpha}_t}$
    Predict $\hat{x}_{0,i}(x_{t,i}) \leftarrow \big(x_{t,i} - \sigma_t \epsilon_\phi(x_{t,i}, t)\big)/\sqrt{\bar{\alpha}_t}$
    if $t \geq T/2$ **then**
        Compute semantic directions $\mathbf{u}_{t,i}^{\text{sem}}$ from τ_i according to Sec. 3.3
        Set $\mathbf{u}_{t,i}^{\text{S2P}} \leftarrow \mathbf{u}_{t,i}^{\text{sem}}$
    else
        Activate layers $\mathcal{A}_t \leftarrow \{L - s_t + 1, \ldots, L\}$
        Compute $\mathcal{L}_{\text{grad}}(\mathbf{X}_t)$ from gradients according to Sec. 3.4
        Compute gradient directions $\mathbf{u}_{t,i}^{\text{grad}}$ according to Sec. 3.4
        Set $\mathbf{u}_{t,i}^{\text{S2P}} \leftarrow \mathbf{u}_{t,i}^{\text{grad}}$
    end if
    Update all $x_{t-1,i} \leftarrow \text{DDIM}(x_{t,i}, \epsilon_\phi(x_{t,i}, t)) - \gamma_t \sigma_t \sqrt{D_x}\, \mathbf{u}_{t,i}^{\text{S2P}}$
end for
return $\{x_{0,i}\}_{i=1}^{B}$

toward both the semantic and fine-grained details encoded in clients' shared gradients, with the goal of improving gradient inversion attacks (GIAs) in federated learning.

A.1.3 PRIVACY–UTILITY–EFFICIENCY TRADE-OFFS

Gradient-level defenses introduce perturbations through mechanisms such as noise injection and orthogonal-subspace sampling (Wei et al., 2020; Zhang et al., 2025a). Image-level defenses introduce privacy-preserving surrogate data to reduce leakage (Fan et al., 2025a; Sun et al., 2021). Although stronger perturbations can improve resistance to reconstruction, they may also reduce model accuracy. Cryptographic defenses introduce secure computation through secure multi-party computation (SMPC) (Lv et al., 2026) or homomorphic encryption (Costa et al., 2026). These approaches avoid directly perturbing the training signal but introduce substantial computational and communication overhead.

A.2 PSEUDO-CODE FOR SEM2PIX

Algorithm 1 summarizes the semantic-to-pixel reconstruction procedure.

A.3 ATTACK COST AND CONVERGENCE

We evaluate the computational cost of SEM2PIX on an NVIDIA RTX 3090 GPU. It requires approximately 20 minutes for 1,000 reverse steps, compared with 15 minutes for each of GGSS-R (Meng et al., 2026) and GGDM (Gu et al., 2024). The additional cost mainly comes from CLIP-based semantic guidance in the first stage. Figure 5 further shows the reconstruction quality of SEM2PIX measured by Gradient Magnitude Similarity Deviation (GMSD) (Xue et al., 2013), where lower values indicate better structural fidelity. In our evaluation, 1,000 reverse steps recover semantic structure and fine-grained details, with only minor perceptual changes after approximately 100 steps.

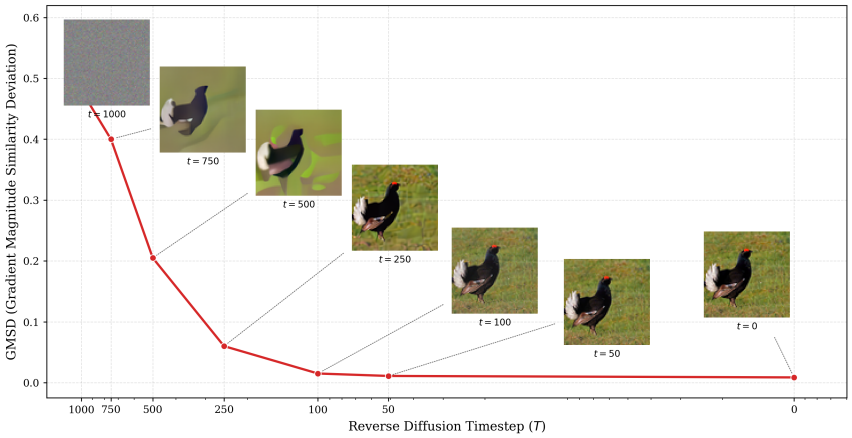
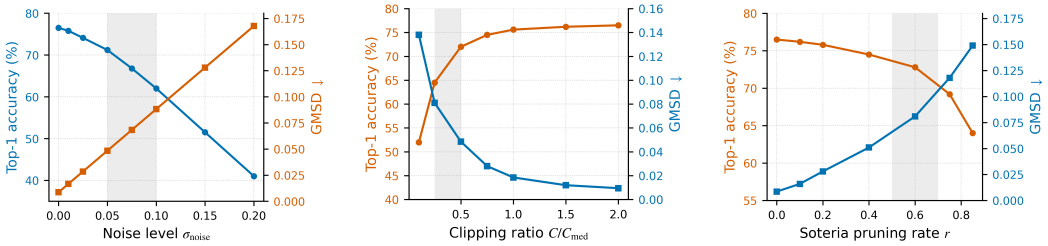


Figure 5: Gradient Magnitude Similarity Deviation (GMSD between reconstructed and ground-truth images over reverse sampling steps.



(a) Noise defense (b) Gradient clipping defense (c) Soteria defense

Figure 6: Trade-off analysis between FL model accuracy (%) and Gradient Magnitude Similarity Deviation (GMSD) across three gradient defenses on ImageNet. The gray-shaded regions highlight the trade-off range for each defense mechanism.

A.4 Evaluation of Utility-Privacy Trade-offs

We examine the trade-off between model utility and privacy with one local training step and a batch size of 1. Since the other attacks perform worse than the proposed method, their results are not reported. Figure 6 illustrates the utility–privacy trade-off between accuracy and GMSD under different defenses on the ImageNet-1K validation set. We partition the images into 100 non-overlapping client subsets. Generally, a GMSD value greater than 0.1 indicates that the reconstructed images no longer retain meaningful information about the original private images. Moreover, our findings underscore the inherent challenge of balancing model utility and privacy preservation in FL. For example, at a noise level of 0.2, model performance decreases from a top-1 accuracy of 76.5% without perturbation to 41.0%.

A.5 Practical FL Evaluation

In this section, we conduct an evaluation designed to reflect real-world federated learning (FL) environments (Carletti et al., 2025). Specifically, we evaluate the robustness of diffusion-based attacks, including GGSS-R (Meng et al., 2026), GGDM (Gu et al., 2024), and the proposed Sem2Pix, on ImageNet.

A.5.1 Impact of Larger Batch Sizes

We evaluate Sem2Pix on ImageNet with batch sizes $B \in \{16, 32, 64\}$ under distinct-label and duplicate-label batch settings. The former consists of batches in which all images have different class labels, whereas the latter consists of batches in which some images share a class label. As

Table 4: Reconstruction performance on ImageNet across batch sizes for distinct-label and duplicate-label client batches.

		Distinct Labels					Duplicate Labels				
Batch Size	Method	LPIPS ↓	SSIM ↑	PSNR ↑	GMSD ↓	CLIP similarity ↑	LPIPS ↓	SSIM ↑	PSNR ↑	GMSD ↓	CLIP similarity ↑
$B = 16$	GGDM (Gu et al., 2024)	0.8950	0.1120	8.950	0.3850	0.1420	0.9210	0.0890	7.820	0.4120	0.1050
	GGSS-R (Meng et al., 2026)	0.6120	0.2150	12.450	0.2240	0.4100	0.6540	0.1820	11.200	0.2520	0.3520
	SEM2PIX	**0.5850**	**0.2410**	**15.820**	**0.2120**	**0.5850**	**0.6210**	**0.2080**	**14.150**	**0.2350**	**0.5120**
$B = 32$	GGDM (Gu et al., 2024)	0.9450	0.0750	7.250	0.4210	0.0980	0.9620	0.0510	6.450	0.4520	0.0650
	GGSS-R (Meng et al., 2026)	0.7250	0.1580	10.120	0.2850	0.2950	0.7680	0.1250	9.240	0.3120	0.2310
	SEM2PIX	**0.6680**	**0.1950**	**13.450**	**0.2450**	**0.4620**	**0.7120**	**0.1620**	**11.950**	**0.2780**	**0.3950**
$B = 64$	GGDM (Gu et al., 2024)	0.9820	0.0420	5.820	0.4650	0.0620	0.9910	0.0210	4.950	0.4920	0.0380
	GGSS-R (Meng et al., 2026)	0.8350	0.1120	8.450	0.3420	0.1850	0.8720	0.0820	7.520	0.3750	0.1320
	SEM2PIX	**0.7520**	**0.1480**	**11.120**	**0.2980**	**0.3450**	**0.7950**	**0.1180**	**9.850**	**0.3320**	**0.2820**

Table 5: Reconstruction performance on ImageNet under multiple local update steps (E).

Local Steps	Method	LPIPS ↓	SSIM ↑	PSNR ↑	GMSD ↓	CLIP similarity ↑
$E = 2$	GGDM (Gu et al., 2024)	0.7820	0.2210	13.5100	0.3420	0.1850
	GGSS-R (Meng et al., 2026)	0.3850	0.3820	17.8200	0.0980	0.7850
	SEM2PIX	**0.2140**	**0.5820**	**21.4500**	**0.0450**	**0.9120**
$E = 4$	GGDM (Gu et al., 2024)	0.8410	0.1580	10.8400	0.3680	0.1520
	GGSS-R (Meng et al., 2026)	0.5210	0.2640	14.1200	0.1720	0.5840
	SEM2PIX	**0.3820**	**0.4150**	**18.2300**	**0.1120**	**0.8140**
$E = 8$	GGDM (Gu et al., 2024)	0.9120	0.0950	8.1200	0.4120	0.1180
	GGSS-R (Meng et al., 2026)	0.6840	0.1720	11.2500	0.2450	0.3920
	SEM2PIX	**0.5410**	**0.2780**	**15.1100**	**0.1850**	**0.6720**

shown in Table 4, reconstruction quality decreases as the batch size increases and deteriorates further when labels are repeated. These results highlight the difficulty of reconstructing individual samples from gradients aggregated over larger batches, particularly when multiple samples belong to the same class. These findings underscore the significant challenges involved in achieving high-fidelity reconstruction in FL with larger batches and duplicate labels, consistent with the challenges discussed in HFGradInv (Ye et al., 2024) and GLAD (Xiang et al., 2026).

A.5.2 Impact of Local Training Steps

To reduce communication costs, clients may perform multiple local training steps before transmitting an update to the server. As the number of local steps E increases, the shared client update accumulates the effects of gradients computed across successive optimization steps rather than corresponding to a gradient computed at a single model state. This weaker correspondence makes it more difficult to recover individual private images. As shown in Table 5, SEM2PIX consistently outperforms the other diffusion-based attacks across the evaluated values of E. We attribute this robustness in part to vision–language semantic guidance (Section 3.3), which establishes class-compatible global structure at the beginning of the reverse process and reduces reliance on an exact gradient-matching signal.

A.5.3 Impact of Communication Rounds

Gradient inversion attacks can be launched at any communication round in FL. However, reconstruction quality is generally higher during the initial communication rounds than during later rounds (Carletti et al., 2025). Moreover, a target client may not participate in the initial rounds. Therefore, rather than evaluating GIAs only at the beginning of training, we assess their performance during the initial, middle, and late stages of FL. Overall, the results in Table 6 characterize the privacy leakage risk across different stages of FL training.

A.5.4 Impact of Deeper Model Architecture

We evaluate SEM2PIX using three target model architectures, namely ResNet-18 (He et al., 2016), DenseNet-121 (Huang et al., 2017), and ViT-B/16 (Dosovitskiy et al., 2021). Table 7 reports the

Table 6: Reconstruction performance on ImageNet across FL communication rounds R.

Stage	Method	LPIPS ↓	SSIM ↑	PSNR ↑	GMSD ↓	CLIP similarity ↑
Initial ($R = 1$)	GGDM (Gu et al., 2024)	0.6991	0.3308	17.1600	0.3016	0.2286
	GGSS-R (Meng et al., 2026)	0.1787	0.5617	22.3760	0.0195	0.9561
	SEM2PIX	**0.0776**	**0.7451**	**25.8470**	**0.0086**	**0.9864**
Mid ($R = 50$)	GGDM (Gu et al., 2024)	0.7950	0.2180	12.6500	0.3380	0.1780
	GGSS-R (Meng et al., 2026)	0.4120	0.3650	16.8400	0.1250	0.7620
	SEM2PIX	**0.2350**	**0.5510**	**20.9100**	**0.0580**	**0.9080**
Late ($R = 100$)	GGDM (Gu et al., 2024)	0.8840	0.1240	9.4500	0.3810	0.1380
	GGSS-R (Meng et al., 2026)	0.5980	0.2280	13.1200	0.2080	0.5120
	SEM2PIX	**0.3980**	**0.3890**	**17.3400**	**0.1280**	**0.8250**

Table 7: Reconstruction performance of SEM2PIX across target model architectures on ImageNet.

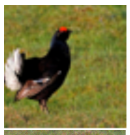
Architecture	Reconstruction	LPIPS ↓	SSIM ↑	PSNR ↑	GMSD ↓	CLIP similarity ↑
ResNet-18 (Baseline)		**0.0776**	**0.7451**	**25.8470**	**0.0086**	**0.9864**
DenseNet-121		0.1425	0.6520	22.1400	0.0185	0.9825
ViT-B/16		0.2150	0.5410	19.8200	0.0294	0.9782

Table 8: Component ablation of SEM2PIX on ImageNet. "Semantic" denotes vision–language semantic guidance. "All-layer" matches all gradients simultaneously, while "Stage-wise" progressively incorporates gradients from later to earlier layers.

Semantic Guidance	Gradient Guidance	LPIPS ↓	SSIM ↑	PSNR ↑	GMSD ↓	CLIP similarity ↑
No	All-layer	0.1950	0.5750	22.6000	0.0280	0.9480
Yes	None	0.5150	0.4150	17.8000	0.2120	0.9100
Yes	All-layer	0.1120	0.6810	24.7000	0.0185	0.9760
Yes	**Stage-wise**	**0.0776**	**0.7451**	**25.8470**	**0.0086**	**0.9864**

reconstruction results. We observe that larger models exacerbate privacy leakage, consistent with the findings of Fan et al. (2025b); Meng et al. (2026).

A.6 ABLATION STUDIES

A.6.1 CONTRIBUTION OF EACH COMPONENT

We evaluate the individual and joint effects of vision–language semantic guidance and stage-wise gradient guidance on ImageNet. Table 8 shows that the use of both mechanisms improves all metrics and that the stage-wise variant achieves the best overall performance. These results confirm that semantic guidance establishes class-compatible global structure, while stage-wise gradient guidance recovers fine-grained details.

A.6.2 ABLATION OF LABEL LEAKAGE

We evaluate the effect of the label-inference strategy on reconstruction quality. Accurate labels are essential because they determine both the target text prompts and the labels used for gradient

Table 9: Ablation of label-inference strategies on ImageNet.

Label Inference	LPIPS ↓	SSIM ↑	PSNR ↑	GMSD ↓	CLIP similarity ↑
iDLG label inference (Zhao et al., 2020)	0.1840	0.6020	21.7300	0.0520	0.9430
Magnitude-aware (Ours)	**0.0776**	**0.7451**	**25.8470**	**0.0086**	**0.9864**

Table 10: Ablation of the guidance strength γ_t on ImageNet.

γ_t	LPIPS ↓	SSIM ↑	PSNR ↑	GMSD ↓	CLIP similarity ↑
0.10	0.1845	0.6218	22.1470	0.0412	0.9581
0.20	0.1127	0.6984	24.5360	0.0179	0.9772
0.30	**0.0776**	**0.7451**	**25.8470**	**0.0086**	**0.9864**
0.40	0.0968	0.7215	25.1260	0.0127	0.9821
0.50	0.1432	0.6679	23.8140	0.0265	0.9704

matching. Table 9 shows that our magnitude-aware strategy is more effective than iDLG (Zhao et al., 2020) across all metrics. This improvement arises because our magnitude-aware score measures the total magnitude of the negative entries in each row of the FC-layer gradient, making label inference less sensitive to isolated sign changes caused by gradient perturbations.

A.6.3 Ablation of Guidance Strength

We evaluate the effect of the guidance strength γ_t on reconstruction quality. Weak guidance may not sufficiently constrain the reverse trajectory, whereas excessive guidance may move it away from the pretrained diffusion prior. Table 10 shows that $\gamma_t = 0.30$ achieves the best performance across all metrics.