PARTIAL IMAGES RAISE FURTHER PRIVACY CONCERNS IN PERSONALIZED DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized diffusion models can adapt to user-specific subjects from a small set of training images, enabling generation in novel scenes, styles, and compositions. Their checkpoints can be shared for reuse without disclosing the private fine-tuning images, which may appear privacy-preserving. However, information retained in the checkpoints can still enable recovery of those images. While prior extraction methods primarily use free-form generation without targeting a particular photograph, partially obscured or redacted versions of private images may also remain accessible through social media or other public sources. These partial observations provide visual cues for reconstructing a specific photograph, yet their privacy implications remain largely unexplored. We therefore investigate how this information can raise further privacy concerns in released personalized checkpoints. To demonstrate this risk, we introduce Partial-Image-Assisted Reconstruction (PIAR), which combines image inpainting with contrastive guidance to reconstruct complete images from partial observations. At each denoising step, contrastive guidance amplifies the difference between the personalized and pretrained models' noise predictions, strengthening the influence of information learned during personalization while remaining conditioned on the visible content. Experiments on CelebA-HQ, DreamBooth, and CustomConcept101 show that PIAR reconstructs private images substantially more accurately than prior free-form extraction. On CelebA-HQ, PIAR achieves an average image similarity of 0.909, compared with 0.537 for the strongest evaluated free-form baseline. The risk extends beyond the fine-tuning images, as PIAR also aids reconstruction of other photographs of the same subject that were not used during personalization. Together, these findings highlight how partial visual information can expose sensitive content through released checkpoints and motivate the need for stronger privacy protections for personalized diffusion models.

1 INTRODUCTION

Diffusion models have become prevalent for generating high-quality images and support a wide range of applications, including personalization (Gal et al., 2023; Ruiz et al., 2023), image editing and inpainting (Meng et al., 2022; Lugmayr et al., 2022), and watermarking (Wen et al., 2023). Among these applications, personalization has attracted particular interest: it enables a general-purpose model to adapt to user-specific subjects from a small set of reference images. This capability is especially useful in text-to-image generation, where natural-language prompts provide flexible control over rendering the learned subject in novel scenes, styles, and compositions beyond its original examples. Such personalization is commonly achieved by fine-tuning a pretrained text-to-image diffusion model (Rombach et al., 2022; Saharia et al., 2022) on the reference images.

Personalized diffusion model checkpoints are often shared through public model repositories and online communities, such as Civitai and Hugging Face (Civitai, 2023; Hugging Face, 2026), so that others can directly use the personalized model without repeating the fine-tuning process or accessing the original reference images. At first glance, sharing only the checkpoint may appear privacy-preserving because the underlying images are not directly disclosed. However, releasing the checkpoint may still expose information about the data used for personalization, as diffusion models are known to memorize and reproduce their training samples (Somepalli et al., 2023b;a). This risk is particularly concerning for personalized models, whose fine-tuning data often consist of user-

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.

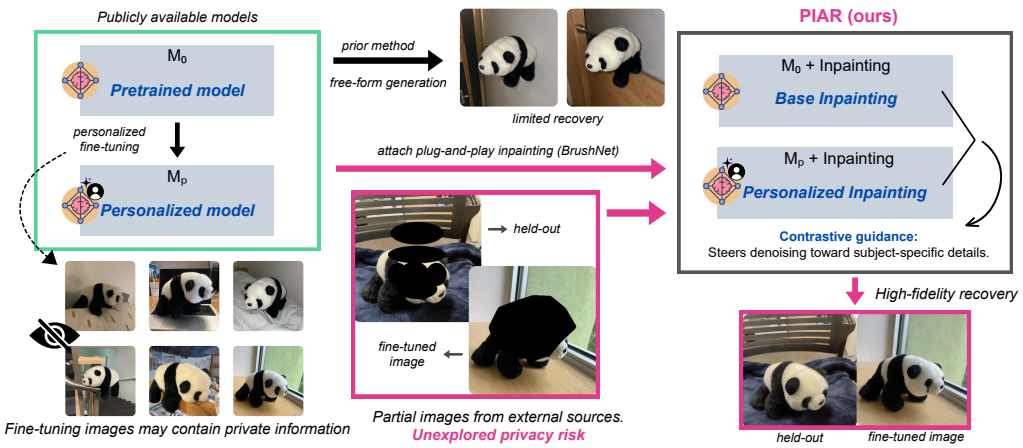


Figure 1: Overview of the privacy setting. Personalized checkpoints can be shared without disclosing their private fine-tuning images, while partial views of the same subject may remain accessible from external sources. PIAR combines image inpainting with contrastive guidance derived from the pretrained and personalized models to reconstruct both fine-tuning and held-out photographs, revealing how partial visual information can amplify privacy leakage beyond free-form extraction.

specific and potentially sensitive content, such as human faces, personal objects, or other identifiable visual information. Recent work has demonstrated this risk by recovering fine-tuning samples from released personalized checkpoints through free-form generation (Wu et al., 2025). While these findings establish that the checkpoint alone can expose fine-tuning data, additional information about the underlying samples may also be available in practice. Consequently, considering the checkpoint in isolation may underestimate the actual privacy risk.

In reality, information related to the fine-tuning samples may also become available from external sources in unexpected ways. For example, images used for personalization may unintentionally remain accessible through social media, public webpages, or street-view services, even when parts of the image have been obscured to protect sensitive information (Orekondy et al., 2018; Bertran et al., 2019). When combined with a personalized diffusion model, such auxiliary visual information can substantially amplify privacy leakage by providing instance-specific context and spatial constraints beyond what is available from the checkpoint alone. However, the privacy implications of such side information remain largely unexplored. This raises the question: **To what extent can auxiliary visual information amplify privacy leakage from personalized diffusion models?**

To better characterize this risk and inform privacy-preserving design, we study a masked-image setting in which a partially observed photograph of the personalized subject specifies the reconstruction target. To demonstrate this risk, we introduce Partial-Image-Assisted Reconstruction (PIAR) illustrated in Figure 1. The visible region narrows the possible completions, while the difference between the pretrained and personalized models provides a personalization signal that guides recovery. Unlike prior training-data extraction approaches that search for untargeted training examples through free-form generation (Carlini et al., 2023), our method instead performs targeted reconstruction of the specified image. On CelebA-HQ, PIAR increases average image similarity from 0.537 to 0.909, and the fraction of successful reconstructions exceeding a strict similarity threshold from 0.144 to 0.966. The risk also extends to photographs not used during personalization: the personalized checkpoint aids reconstruction of held-out images of the same subject.

Our contributions are three-fold as follows. We identify masked fine-tuning images as an important source of auxiliary information for studying privacy leakage from personalized diffusion models, revealing risks beyond those posed by the released checkpoint alone. We develop a masked-image reconstruction method to study this privacy risk, demonstrating substantially stronger recovery than prior free-form extraction. Through systematic evaluation across datasets, masking conditions, and reconstruction budgets, we characterize when and to what extent this auxiliary visual information

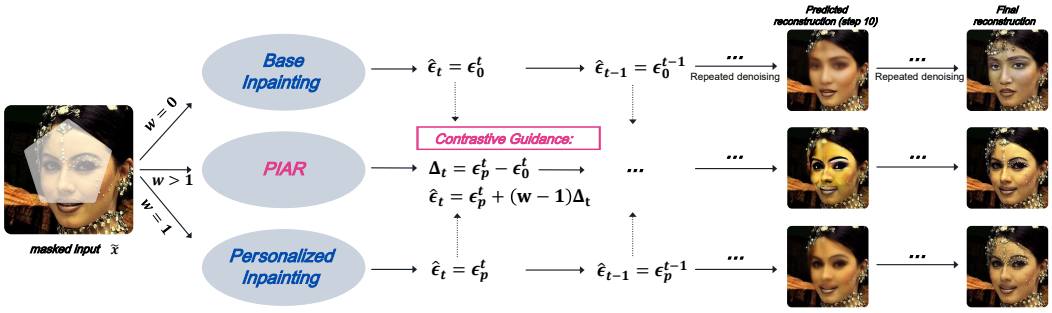


Figure 2: PIAR's contrastive guidance during denoising. Guidance weights $w = 0$, 1, and > 1 yield Base Inpainting, Personalized Inpainting, and PIAR with $w = 3$ used in our experiments, respectively. Each method follows its own denoising trajectory; within PIAR, both models evaluate the same current latent z_t and masked input, and their prediction difference Δ_t guides reconstruction. We show the predicted clean image $\hat{x}_0$ at the 10th denoising step and the final reconstruction after 50 steps. PIAR more faithfully recovers the ornament on the target's forehead than Personalized Inpainting, illustrating how contrastive guidance helps recover local details. The mask is translucent for illustration but blacked out during inference.

amplifies privacy leakage. Together, these findings highlight a previously overlooked leakage pathway and motivate the development of stronger privacy-preserving defenses.

2 Background and Related Work

Personalized Diffusion Models. Representative personalization methods differ in how they incorporate new concepts into pretrained diffusion models. DreamBooth fine-tunes the diffusion model weights to associate a unique identifier with a target subject (Ruiz et al., 2023), while LoRA enables parameter-efficient personalization by learning low-rank updates (Hu et al., 2022). Custom Diffusion instead updates only the key and value projections in cross-attention layers with a modifier token, enabling efficient personalization and concept composition (Kumari et al., 2023). More recently, APT introduces adaptive training and regularization to mitigate overfitting and preserve pretrained knowledge during personalization with limited reference data (Chae et al., 2025). Together, these methods have made subject-specific generation increasingly practical. In this work, we use DreamBooth-personalized models to investigate the privacy risks arising from partially observed images of the personalized subject.

Diffusion-Based Image Inpainting. Image inpainting reconstructs missing regions of an image conditioned on the visible content. Diffusion-based approaches have enabled high-quality inpainting through different forms of image conditioning, including sampling-based methods such as RePaint (Lugmayr et al., 2022), latent-space blending (Avrahami et al., 2023), and dedicated inpainting models such as PowerPaint (Zhuang et al., 2024) and BrushNet (Ju et al., 2024). More recently, VIPaint (Agarwal et al., 2026) uses hierarchical variational inference to improve inpainting with pretrained diffusion models. While these methods primarily aim to generate plausible and visually consistent completions, they are not designed to recover private training samples or expose information memorized by the underlying model. In PIAR, inpainting guides reconstruction of the missing content, while memorization in the personalized model provides the signal needed to recover the hidden training content. As shown in Section 5.3, inpainting alone is insufficient: without the relevant memorized information, reconstruction quality clearly drops for the target image.

Training Data Extraction from Diffusion Models. Diffusion models have been shown to memorize and reproduce individual training samples, raising concerns about privacy and copyright. Carlini et al. (2023) demonstrated that training images can be extracted from large diffusion models such as Stable Diffusion and Imagen using a generate-and-filter pipeline. More recently, SIDE (Chen et al., 2026) introduces surrogate conditions derived from clusters of model-generated images to steer generation toward memorized samples, enabling more targeted extraction within a free-form generation

setting. Its guidance targets these clusters without using the visible regions of a particular target photograph. In the context of personalized diffusion models, the closest work to ours is FineXtract (Wu et al., 2025), which extracts fine-tuning samples through free-form generation. Unlike SIDE and FineXtract, we study a setting where a partially observed image specifies the target in advance and turns extraction into reconstruction of a particular training image.

3 Problem Formulation and Threat Model

Let M_0 denote a pretrained text-to-image diffusion model and M_p its personalized counterpart, fine-tuned on a personalized subject. We consider a target photograph $x^\star$ of a personalized subject and a partially observed version $\tilde{x}$ obtained by masking or redacting part of $x^\star$. We assume access to a publicly available personalized model M_p, its corresponding pretrained model M_0, and a partially observed target image $\tilde{x}$ obtained from an external source. The attacker's goal is to reconstruct the target image $x^\star$ given this information. Unlike free-form extraction, which searches for likely training samples without conditioning on a partially observed photograph, our setting focuses on reconstructing a designated image, as illustrated in Figure 1. While our primary focus is on photographs used during fine-tuning, we also examine held-out photographs of the same subjects.

This setting examines the privacy risk of releasing personalized checkpoints under the possibility of masked images, as well as whether obscuring a photograph provides sufficient protection when a personalized checkpoint of the same subject is also publicly available. We require neither the original fine-tuning prompt nor the subject's special token, since access to the partially observed photograph does not necessarily lead to this information. Our prompt ablation in Appendix C.2 compares reconstruction with and without the subject's special token, showing that substantial recovery is possible even without it. Further combining with prompts describing the targeted image can lead to even better reconstruction, which we leave as an important future direction due to controlled experimental design.

4 Masked-Image Reconstruction with PIAR

We propose Partial-Image-Assisted Reconstruction (PIAR), a method that reconstructs a masked target image by combining diffusion-based inpainting with the information introduced during personalization. We attach the same pretrained BrushNet inpainting module (Ju et al., 2024) to both M_0 and M_p, forming two branches that we refer to as *Base Inpainting* and *Personalized Inpainting*, respectively. Both branches receive the same noisy latent, masked image, mask, and an empty text prompt (Figure 2). BrushNet incorporates the visible-image constraint into the denoising process, while the personalized model retains information learned from the fine-tuning data. Since the two branches share the same conditioning and differ only in whether the underlying diffusion model has been personalized, comparing their denoising predictions captures how personalization changes the reconstruction under the same visual context. We use this signal as our *contrastive guidance* described below.

4.1 Contrastive Guidance

At each denoising step t, both inpainting branches predict noise from the current noisy latent z_t, conditioned on the masked image $\tilde{x}$, mask m, and text prompt c. We denote these predictions by $\epsilon_0^t \equiv \epsilon_0(z_t, t, \tilde{x}, m, c)$ and $\epsilon_p^t \equiv \epsilon_p(z_t, t, \tilde{x}, m, c)$ for Base Inpainting and Personalized Inpainting, respectively, with c set to an empty prompt. We define their difference as

$$\Delta_t = \epsilon_p^t - \epsilon_0^t. \tag{1}$$

Since the two branches are evaluated under identical conditions, Δ_t represents the change in denoising prediction induced by personalization. Motivated by classifier-free guidance (CFG) (Ho & Salimans, 2022) and following the cross-model guidance introduced by FineXtract (Wu et al., 2025), we use this difference as the guidance direction and modify the personalized model prediction as

$$\hat{\epsilon}_t = \epsilon_p^t + (w-1)\Delta_t. \tag{2}$$

Here, w controls the strength of contrastive guidance. When $w = 0$, the formulation reduces to Base Inpainting. When $w = 1$, it reduces to Personalized Inpainting, which already uses information

learned during fine-tuning but does not necessarily reproduce the target's specific details. Using $w > 1$ gives greater weight to the change introduced by personalization, encouraging reconstruction to follow this signal more strongly while remaining conditioned on the visible context. We use $w = 3$ throughout all experiments. An empirical ablation study is reported in Appendix C.1.

4.2 Reconstruction and Candidate Selection

Starting from a noisy latent, we iteratively denoise using $\hat{\epsilon}_t$ while conditioning on the partially observed image and mask. For each masked input, we generate a small set of N stochastic reconstructions. We extract features using SSCD (Self-Supervised Descriptor for Image Copy Detection) (Pizzi et al., 2022), a pretrained image descriptor designed for copy detection, and select the output whose feature is closest to the mean feature of the set. The target image $x^\star$ is not involved in this selection and is used only for evaluation.

Free-form extraction methods seek to recover training images without conditioning on a targeted partial image, and involve large pools of generated candidates and clustering at different stages. In contrast, masked-input reconstruction targets a particular image, and we select a representative directly without needing clustering. As shown in Figure 5(b), PIAR already achieves strong reconstruction with $N = 1$, demonstrating its effectiveness without requiring a large generation pool.

5 Experiments

5.1 Experimental Setup

Datasets and personalized models. We conduct our main experiments on CelebA-HQ (Karras et al., 2018), the DreamBooth dataset (Ruiz et al., 2023), and CustomConcept101 (Kumari et al., 2023), covering face identities and object subjects. Following FineXtract (Wu et al., 2025) for fair comparisons, we use Stable Diffusion v1.4 (Rombach et al., 2022) as the pretrained model and DreamBooth (Ruiz et al., 2023) for personalization.

We personalize a separate checkpoint for each of 50 subjects from CelebA-HQ and 30 subjects from each object dataset. Each checkpoint is personalized using six reference images for CelebA-HQ and CustomConcept101, and four to six reference images for the DreamBooth dataset.

Both masked-input and free-form methods are evaluated on all fine-tuning photographs of each subject. For masked-input methods, each photograph is paired with one of ten small masks (32.3% average coverage), assigned by a fixed rotation. Base Inpainting, Personalized Inpainting, and PIAR use the same masked inputs. Dataset sources, subject selection, and preprocessing are detailed in Appendix A.1, and personalization settings in Appendix A.2.

Baselines. We compare PIAR against four baselines: two free-form baselines, classifier-free guidance (CFG) (Ho & Salimans, 2022) and FineXtract (Wu et al., 2025), and two masked-input baselines, *Base Inpainting* and *Personalized Inpainting*. Following FineXtract's default setting, the free-form baselines receive the original training prompt, including the subject's special token. In our standard DreamBooth setup, the same prompt is used to fine-tune on all reference images of that subject. The masked-input methods use the same partially observed target photograph and an empty prompt, requiring no knowledge of the training prompt or special token. These comparisons assess recovery under free-form extraction and masked-input reconstruction, with the two inpainting baselines isolating the contributions of personalization and PIAR's guidance.

Generation and candidate selection. For each masked input, we generate $N = 20$ candidates and select their SSCD medoid as the representative reconstruction, as described in Section 4. Each representative is evaluated against its corresponding target photograph.

Following FineXtract (Wu et al., 2025), we generate $50 \times N_0$ images per checkpoint for each free-form baseline, where N_0 is the number of fine-tuning images. We apply SSCD-based clustering to form an extracted set of N_0 representatives without accessing the ground-truth images; details are provided in Appendix A.4.

For evaluation, each fine-tuning photograph is matched separately for each metric to the extracted representative yielding the best score. This gives free-form methods the benefit of selecting the best match for each target under each metric, whereas each masked-input reconstruction is evaluated only against its designated target. Scores are averaged over all fine-tuning photographs within each subject. As we will see, even under such an evaluation protocol that favors free-form methods, PIAR still provides more accurate target image reconstruction.

Evaluation metrics. We use cosine similarity between SSCD features to measure image similarity and report its average as Average Similarity (AS). We additionally report Average Extraction Success Rate (A-ESR@τ) (Wu et al., 2025), defined as the mean subject-level fraction of photographs whose SSCD similarity exceeds τ. We report $\tau = 0.8$ in the main comparisons and also present other tested τ values in Appendix B. For CelebA-HQ, we additionally measure facial identity similarity using ArcFace embedding cosine similarity (Deng et al., 2019). We also report CLIP image-embedding cosine similarity (Radford et al., 2021), LPIPS (Zhang et al., 2018), and SSIM (Wang et al., 2004) to assess semantic, perceptual, and structural similarity, respectively. All metrics are computed over the full image.

Implementation and aggregation. DreamBooth models are trained for 200 optimization steps per fine-tuning image, so the total training budget scales with the amount of personalization data. For both method groups, scores are first averaged over all fine-tuning photographs within each subject, then across subjects within each dataset. The main comparison tables report the mean and sample standard deviation of these dataset-level scores across five independent runs, each with new training and sampling seeds. Subject selections, image splits, and mask assignments remain fixed across runs. The sweeps and held-out evaluation use a single run, with aggregation details provided in Appendix A.5. Additional experimental details are provided in Appendix A.

5.2 Main Results

Table 1 compares PIAR with free-form extraction and masked-input baselines on CelebA-HQ, DreamBooth, and CustomConcept101. PIAR substantially outperforms the free-form baselines across all three datasets. For example, on CelebA-HQ, PIAR achieves an AS of 0.909 compared with 0.537 for FineXtract, and an A-ESR@0.8 of 0.966 compared with 0.144. Although reconstruction is generally less accurate on the object datasets, PIAR retains its advantage over the baselines. Figure 3 provides qualitative examples from all three datasets, highlighting the recovery of the target's appearance and local details. Additional reconstruction examples are provided in Appendix B.6.

The masked-input baselines further show that this improvement cannot be explained by inpainting alone. On CelebA-HQ, Base Inpainting achieves an AS of 0.658, while replacing the pretrained model with the personalized checkpoint increases AS to 0.886 and A-ESR@0.8 from 0.071 to 0.919. PIAR further improves these results to 0.909 AS and 0.966 A-ESR@0.8. The same pattern is observed on the object datasets. These results show that partial visual information becomes considerably more revealing when combined with information encoded in the personalized checkpoint, while contrastive guidance provides an additional improvement in reconstruction quality. Appendix B.1 provides the full metrics and variation across subjects from a single run.

5.3 Checkpoint Specificity

We next examine whether a masked input still enables improved reconstruction when paired with a personalized checkpoint that does not contain the target concept. For the target-absent condition, we use a checkpoint from another concept, whose fine-tuning data exclude the target subject. The target image, masks, generation budget, and initial sampling noise remain fixed between the target-present and target-absent conditions. For CelebA-HQ, we evaluate 50 subjects and report the mean and sample standard deviation across five independent runs.

As shown in Table 2, reconstruction quality on CelebA-HQ drops sharply when the target identity is absent from the checkpoint. For Personalized Inpainting, AS falls from 0.886 to 0.647, while PIAR decreases from 0.909 to 0.609. Moreover, target-absent PIAR falls below Base Inpainting in AS (0.609 vs. 0.658), indicating that an unrelated personalized checkpoint provides no average AS advantage over Base Inpainting in this setting. The same pattern holds on DreamBooth and

Table 1: Main reconstruction results on CelebA-HQ, DreamBooth, and CustomConcept101. Personalized Inpainting substantially improves recovery over Base Inpainting, while PIAR further improves reconstruction through contrastive guidance. Metrics are scaled by 100; ± denotes sample standard deviation across five independent runs. ArcFace is evaluated only on CelebA-HQ.

Dataset	Method	AS ↑	A-ESR@0.8 ↑	CLIP ↑	LPIPS ↓	SSIM ↑	ArcFace ↑
CelebA-HQ	CFG	43.8±0.8	1.4±0.6	86.6±0.3	46.5±0.4	40.7±0.4	69.7±0.4
	FineXtract	53.7±0.5	14.4±1.8	89.6±0.2	41.7±0.5	40.8±0.7	77.5±0.2
	Base Inpainting	65.8±0.1	7.1±0.5	81.0±0.1	19.6±0.1	69.0±0.2	18.9±0.3
	Personalized Inpainting	88.6±0.1	91.9±0.8	95.9±0.1	9.9±0.0	77.3±0.1	82.2±0.1
	PIAR	**90.9±0.2**	**96.6±0.7**	**96.4±0.1**	**8.5±0.1**	**78.7±0.1**	**87.3±0.2**
DreamBooth	CFG	45.2±0.8	3.4±0.4	91.3±0.3	56.1±0.9	32.4±0.7	–
	FineXtract	49.9±1.4	7.9±1.4	91.9±0.3	54.0±0.8	30.9±0.8	–
	Base Inpainting	47.7±0.3	0.1±0.2	80.2±0.2	30.1±0.2	54.5±0.1	–
	Personalized Inpainting	68.1±0.2	29.2±1.7	90.1±0.2	21.5±0.1	59.2±0.2	–
	PIAR	**76.4±0.3**	**56.4±1.3**	**92.5±0.3**	**17.4±0.3**	**62.8±0.2**	–
CustomConcept101	CFG	41.7±0.9	0.9±0.5	90.5±0.2	56.8±0.7	29.5±0.7	–
	FineXtract	45.7±0.4	4.6±1.1	91.1±0.2	56.0±0.7	26.9±0.6	–
	Base Inpainting	54.4±0.2	0.8±0.3	86.2±0.3	26.5±0.2	58.3±0.1	–
	Personalized Inpainting	72.5±0.3	38.9±2.1	93.1±0.1	19.1±0.1	62.4±0.1	–
	PIAR	**78.0±0.2**	**55.4±1.3**	**94.4±0.1**	**16.5±0.1**	**64.5±0.1**	–

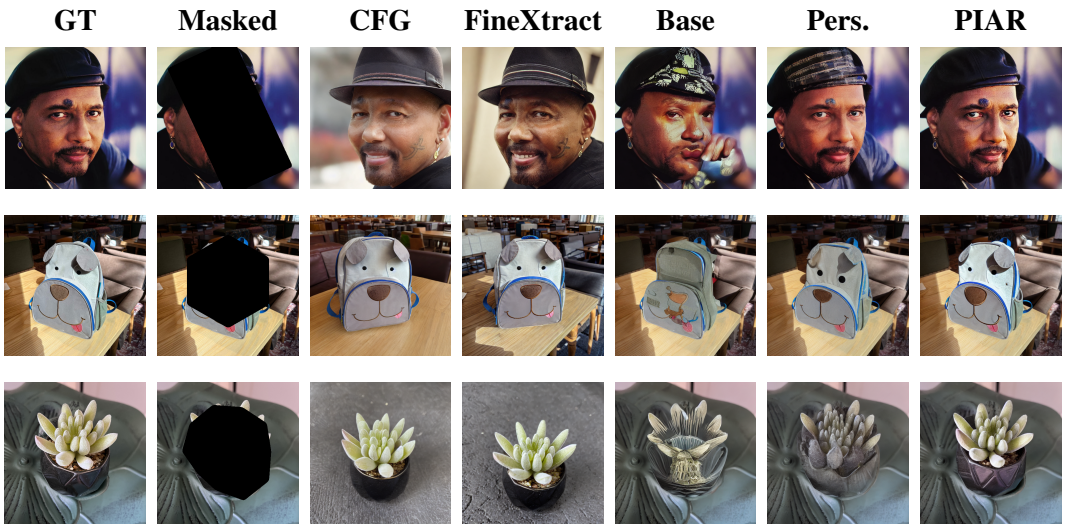


Figure 3: Qualitative comparison of free-form extraction and masked-input reconstruction on CelebA-HQ, DreamBooth, and CustomConcept101 (top to bottom). Each row uses a different mask shape from the small-mask tier. GT and Masked denote the ground-truth image and masked input; Base and Pers. denote Base Inpainting and Personalized Inpainting, respectively. Personalized Inpainting improves recovery over free-form extraction and Base Inpainting, while PIAR further improves reconstruction of the target's appearance and local details.

CustomConcept101, where target-absent PIAR also performs at or below Base Inpainting in AS; full results are provided in Appendix B.2. Together, these results show that the reconstruction advantage depends on target-specific information encoded during personalization, rather than on the masked input and inpainting prior alone.

5.4 Effect of Partial Observation

Mask coverage. We first examine how reconstruction changes as less of the target image remains visible. Figure 8 in Appendix B.3 shows qualitative examples on CelebA-HQ under increasingly large masks, while Figure 5(a) reports reconstruction performance as a function of mask coverage on CelebA-HQ and DreamBooth. For fine-tuning images, Base Inpainting degrades sharply as more of the image is concealed, while reconstruction using the personalized checkpoint remains substantially

Table 2: Checkpoint-specificity results on CelebA-HQ with the target identity either present in or absent from the personalized checkpoint's fine-tuning data. Reconstruction performance drops sharply when the target identity is absent, indicating that the improvement depends on target-specific information encoded during personalization rather than on inpainting alone. Metrics are scaled by 100; ± denotes sample standard deviation across five independent runs.

Method	Target Identity	AS ↑	A-ESR@0.8 ↑	CLIP ↑	LPIPS ↓	SSIM ↑	ArcFace ↑
Base Inpainting	–	65.8±0.1	7.1±0.5	81.0±0.1	19.6±0.1	69.0±0.2	18.9±0.3
Personalized Inpainting	Present	88.6±0.1	91.9±0.8	95.9±0.1	9.9±0.0	77.3±0.1	82.2±0.1
	Absent	64.7±0.1	4.9±1.0	79.9±0.3	17.8±0.1	70.9±0.1	18.6±0.4
PIAR	Present	90.9±0.2	96.6±0.7	96.4±0.1	8.5±0.1	78.7±0.1	87.3±0.2
	Absent	60.9±0.2	1.3±0.5	76.9±0.2	18.6±0.1	70.6±0.1	15.0±0.2

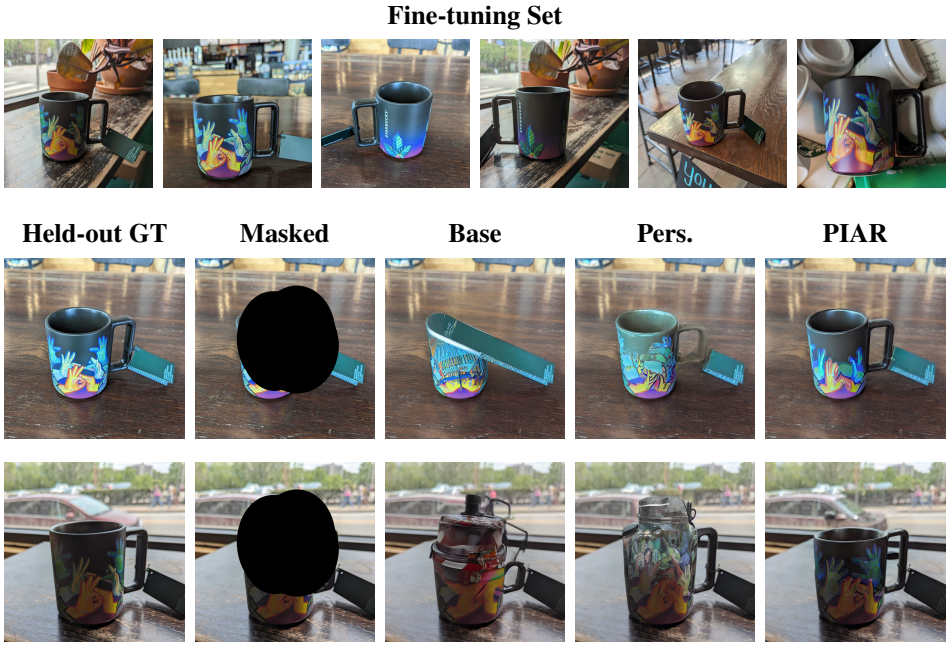


Figure 4: Reconstruction of held-out photographs of the same subject. The top row shows fine-tuning images; the bottom rows show two held-out targets reconstructed under small masks. GT denotes ground truth; Base and Pers. denote Base Inpainting and Personalized Inpainting. PIAR improves recovery over all baselines even when the targets were not used during personalization.

stronger. PIAR consistently achieves the highest AS, and its advantage over Personalized Inpainting grows with mask coverage across all three datasets. On CelebA-HQ, even at 80.1% average mask coverage, PIAR maintains an AS of 0.775, compared with 0.631 for Personalized Inpainting and 0.208 for Base Inpainting. Recovery is more limited on the object datasets under extensive masking, although PIAR continues to outperform both baselines. These results show that contrastive guidance provides larger AS gains as less visual information is available, and that reconstruction remains possible even when much of the target image is concealed. Full results across all mask-coverage levels are provided in Appendix B.3.

Held-out photographs. We next examine whether this effect is limited to photographs that were directly used during personalization. For quantitative evaluation, we repeat the mask-coverage sweep on one held-out photograph per subject, which was excluded from the corresponding checkpoint's fine-tuning set. As shown in Figure 5(a) and 7(a), both Personalized Inpainting and PIAR achieve higher AS than Base Inpainting across all mask tiers on all three datasets. Figure 4 further illustrates how personalization helps recover the appearance and local details of even an unseen photograph. These results show that personalization can aid reconstruction even when the exact pho-

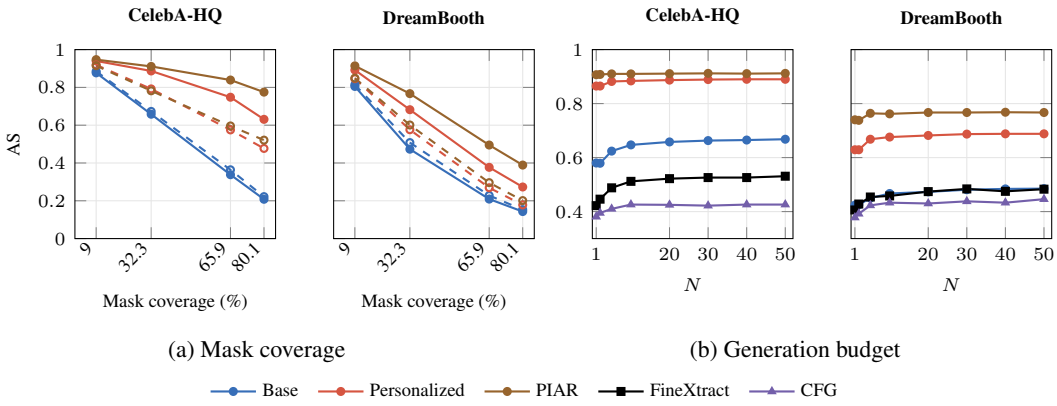


Figure 5: Effect of mask coverage and generation budget on reconstruction. (a) Solid and dashed lines show fine-tuning and held-out photographs, respectively. PIAR improves reconstruction in both settings, particularly at higher mask coverage. (b) PIAR achieves strong reconstruction with a single candidate and quickly saturates, while CFG and FineXtract benefit more from additional generations. N denotes generations per masked input or per fine-tuning image for masked-input and free-form methods, respectively. Results for all three datasets are provided in Figure 7.

tograph was not observed during fine-tuning. Full held-out results across all mask-coverage levels are provided in Appendix B.4.

5.5 Generation Budget

We examine how reconstruction performance changes with the generation budget. For masked-input methods, N denotes the number of candidates generated per masked input. For CFG and FineXtract, N denotes the number of generations per fine-tuning image, giving a total of $N \times N_0$ generations for a checkpoint trained on N_0 images. Thus, evaluating all N_0 fine-tuning photographs uses $N \times N_0$ generations per checkpoint for either method group.

As shown in Figure 5(b) and 7(b), PIAR achieves strong reconstruction with a single candidate across all three datasets, reaching AS values of 0.907, 0.74, and 0.754 on CelebA-HQ, DreamBooth, and CustomConcept101, respectively. Increasing PIAR's budget to $N = 50$ yields only modest gains, highlighting its effectiveness even with very few candidates, while CFG and FineXtract benefit more from additional generations. Full results across all generation budgets are provided in Appendix B.5.

6 Conclusions and Future Directions

Our findings highlight the importance of considering personalized checkpoints and auxiliary visual information together when assessing privacy. Keeping the original photographs private may offer insufficient protection when partially observed versions remain accessible. The held-out results further suggest that personalization can aid reconstruction beyond exact fine-tuning photographs, making the privacy concern relevant to other images of the same subject.

Our evaluation covers three datasets but focuses on DreamBooth personalization with Stable Diffusion v1.4. Future work could examine other personalization methods, model architectures, and forms of visual obfuscation, such as blurring and mosaicing. Our free-form comparisons use the original training prompts, following FineXtract's setting; whether photograph-specific prompts derived from visible regions improve extraction remains an open question. Notably, since PIAR does not use prompt information in our current work, such an informative prompt may also be combined for better reconstruction quality. Through unveiling the privacy risk of personalized diffusion models with masked images as auxiliary information, we hope our work motivates more study toward defense mechanisms for this reconstruction threat.

AI USE STATEMENT

The research motivation and core method originated with the authors. We used generative AI tools to assist with experimental design, method implementation, and interpretation of results, as well as literature search and manuscript preparation. The authors reviewed AI-assisted text for accuracy and consistency with the intended claims, and checked AI-assisted code against the intended methodology. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies a privacy risk of personalized diffusion models. Fine-tuning images are often assumed to stay private once only the checkpoint is shared. We show that this assumption breaks down when a masked or redacted version of an image is also available, for example through social media or other public sources. In that case, the fine-tuning images can be reconstructed with high fidelity. Held-out photographs of the same subject, which were never used for training, also become easier to recover. PIAR could in principle be applied to existing public checkpoints. That is not our goal. We want model creators, platform maintainers, and researchers to take auxiliary visual information into account when assessing the privacy of shared checkpoints, and we want to encourage work on defenses against this kind of reconstruction.

We took several steps to limit potential harm. All experiments use public datasets (CelebA-HQ, the DreamBooth dataset, and CustomConcept101) and public models (Stable Diffusion v1.4, DreamBooth, and BrushNet). We did not collect new images of private individuals. We did not attack any personalized checkpoint shared by third parties. No IRB review was required since no new human-subject data was collected. The only face images in our study come from CelebA-HQ, a standard public benchmark of publicly known individuals, and we use them only to measure reconstruction accuracy. We will release our experiment code so that others can reproduce our results and evaluate defenses. We believe that making this risk visible to the community outweighs the small additional capability our method provides.

REPRODUCIBILITY STATEMENT

Section 4 describes PIAR and its candidate-selection procedure. Appendix A provides dataset preparation, model training, mask construction, sampling settings, and evaluation and aggregation procedures. Appendices B and C report extended results and ablations. Our anonymized code, including experimental configurations, subject and image lists, and masks, is available at `https://anonymous.4open.science/r/PIAR-ICLR/`.

REFERENCES

Sakshi Agarwal, Gabriel Hope, Jimin Heo, and Erik B. Sudderth. VIPaint: Image inpainting with pre-trained diffusion models via variational inference. In *International Conference on Artificial Intelligence and Statistics*, 2026. URL `https://arxiv.org/abs/2411.18929`.

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics*, 42(4), 2023. doi: 10.1145/3592450.

Martin Bertran, Natalia Martinez, Afroditi Papadaki, Qiang Qiu, Miguel Rodrigues, Galen Reeves, and Guillermo Sapiro. Adversarially learned representations for information obfuscation and inference. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 614–623. PMLR, 2019. URL `https://proceedings.mlr.press/v97/bertran19a.html`.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270. USENIX Association, 2023.

JungWoo Chae, Jiyoon Kim, JaeWoong Choi, Kyungyul Kim, and Sangheum Hwang. APT: Adaptive personalized training for diffusion models with limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28619–28628, 2025.

Yunhao Chen, Shujie Wang, Difan Zou, and Xingjun Ma. Side: Surrogate conditional data extraction from diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 128–136, 2026. doi: 10.1609/aaai.v40i1.36972.

Civitai. Civitai: How to use models. `https://github.com/civitai/civitai/wiki/How-to-use-models`, 2023. Accessed September 2026.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL `https://arxiv.org/abs/2207.12598`.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Hugging Face. Hugging face hub: Models. `https://huggingface.co/docs/hub/models`, 2026. Accessed 2026.

Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, 2024.

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*, 2018.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8466–8475, 2018.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494, 2022.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47783–47803, 2023a.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048–6058, 2023b.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. In *Advances in Neural Information Processing Systems*, 2023.

Xiaoyu Wu, Jiaru Zhang, and Steven Wu. Leveraging model guidance to extract training data from personalized diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68060–68077. PMLR, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *European Conference on Computer Vision*, pp. 195–211. Springer, 2024.

A Additional Experimental Details

A.1 Dataset Sources and Preparation

CelebA-HQ. We use the CelebA-HQ images (Karras et al., 2018) as distributed in CelebAMask-HQ (Lee et al., 2020): 30,000 face images at 1024×1024 resolution derived from CelebA (Liu et al., 2015). Identity labels are taken from CelebA's identity annotations. For each identity, we further filter its photographs, keeping those that are pairwise similar (ArcFace cosine similarity ≥ 0.4), and remove near-duplicates using SSCD (similarity > 0.6). We select 50 identities, each with six photographs for fine-tuning and one additional photograph reserved for held-out evaluation. Images are used at their original alignment and are only resized to 512×512; no further cropping or alignment is applied.

DreamBooth Objects. We use the subject dataset released with DreamBooth (Ruiz et al., 2023) (CC-BY-4.0), which is also the object benchmark used by FineXtract (Wu et al., 2025). It contains 30 subjects (21 objects and 9 live subjects) across 15 classes, with 4–6 images per subject. We use all images of each subject for fine-tuning, and use the dataset-provided class noun in the instance prompt (e.g., "a photo of sks1 teapot"). For held-out evaluation, we train a separate checkpoint for each subject, withholding the image with the lowest mean SSCD similarity to the remaining images. All images are square and are resized to 512×512.

CustomConcept101. We use CustomConcept101 (Kumari et al., 2023), which contains 101 concepts with 3–15 images each across 16 categories. We consider concepts with at least seven images and select at most three concepts per category, yielding 30 concepts across 13 categories. For each concept, we use six images for fine-tuning. Among the remaining images, we select the one with the lowest mean SSCD similarity to the fine-tuning images as the held-out photograph. We use the dataset-provided class noun in the instance prompt. Non-square images are center-cropped to a square, and all images are resized to 512×512.

The selected subjects and fine-tuning and held-out image lists for all three datasets are included in our released code.

A.2 Victim Model Training

We personalize Stable Diffusion v1.4 with DreamBooth using the training script provided by Diffusers, following the main fine-tuning settings of FineXtract (Wu et al., 2025). Only the U-Net is fine-tuned; the text encoder and VAE remain frozen. Training runs for $200 \times N_0$ steps at a constant learning rate of 2×10^{-6} with batch size 1, where N_0 is the number of fine-tuning images. The prior-preservation loss is disabled. We use 8-bit Adam, fp16 mixed precision, and gradient checkpointing. Each checkpoint is personalized on a single subject with the instance prompt "a photo of sks1 [class]", where [class] is "person" for CelebA-HQ and the dataset-provided class noun for DreamBooth and CustomConcept101.

A.3 Masks

We use four mask tiers, partial, small, large, and extra-large, each containing ten binary masks at 512×512 resolution with average coverage of 9.0%, 32.3%, 65.9%, and 80.1% of the image, respectively. The small, large, and extra-large tiers share ten shapes of varied geometry, such as polygons, bands, and irregular blobs, centred on the image; the large and extra-large masks are the same shapes scaled up so that their coverage spans 52–80% and 70–90%. The partial tier consists of small, localized masks with varied shapes and positions near the image centre, including combinations of multiple masked regions. Coverage is measured over the full image, and the same masks are used for all datasets, methods, and conditions. In the main experiments, each fine-tuning photograph is paired with one mask of the tier, assigned by a fixed rotation so that all ten masks are used approximately equally, and the same assignment is kept across tiers. The held-out experiment applies all ten masks to each held-out photograph. Figure 6 shows all four mask tiers.

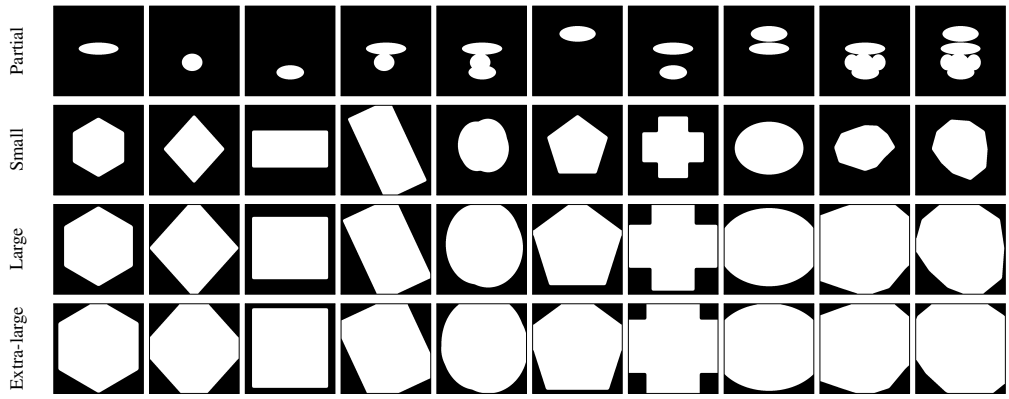


Figure 6: Masks used in our experiments, with ten masks per tier. Average coverage is 9.0%, 32.3%, 65.9%, and 80.1% for the partial, small, large, and extra-large tiers, respectively.

A.4 Generation and Candidate Selection

Masked-input methods. Base Inpainting, Personalized Inpainting, and PIAR use $w = 0$, 1, and 3, respectively, in Eq. (2). All three use the public random-mask BrushNet checkpoint (Ju et al., 2024), DDIM sampling with 50 steps, and an empty prompt. For the main comparisons, we generate $N = 20$ candidates per masked input and select their SSCD medoid. Corresponding candidates share identical initial noise across all three methods and the target-present and target-absent conditions; sampling seeds vary across runs. For the generation-budget sweep, budget n uses the first n candidates from the same pool.

Free-form methods. CFG and FineXtract use the checkpoint's training prompt, with empty prompts for the unconditional branches. Both use DDIM sampling with 50 steps. We set the CFG scale to 3 and use FineXtract's published settings $w' = 3$ and $k = -0.02$. Following FineXtract (Wu et al., 2025), the main comparisons use $50N_0$ generations per checkpoint: 300 for CelebA-HQ and CustomConcept101, and 200–300 for DreamBooth. Each run uses an independently retrained checkpoint and new sampling seeds, with identical initial noise shared between CFG and FineXtract for corresponding candidates.

Free-form candidate selection. Both free-form baselines use the same SSCD-based clique selection, adapted from FineXtract (Wu et al., 2025), we use a default threshold of 0.5. We connect generated images whose pairwise SSCD similarity meets the threshold, then greedily extract disjoint cliques by repeatedly removing the largest maximal clique until no clique of at least two images remains. We retain the N_0 largest cliques and select each clique's SSCD medoid as an extracted image.

For small generation budgets in Section 5.5, fewer than N_0 cliques may be found; we then relax the clustering threshold and allow singleton cliques as needed.

A.5 Evaluation and Aggregation

Metrics. We compute SSCD cosine similarity (Pizzi et al., 2022) between each reconstruction and its target photograph and report the average as AS. We use the SSCD disc-large model with FineXtract's preprocessing: resizing to 256 pixels followed by a 224×224 centre crop. A-ESR@τ measures the fraction of evaluated photographs whose similarity exceeds τ, for $\tau = 0.7$ and 0.8. LPIPS uses the AlexNet backbone, and SSIM is computed over RGB channels, both at 512×512. CLIP similarity uses cosine similarity between image embeddings from OpenAI's ViT-B/32 checkpoint with its default preprocessing.

For CelebA-HQ, we additionally report ArcFace cosine similarity using the InsightFace `buffalo_l` models. Only pairs with a face detected in both images contribute to the ArcFace average. All metrics compare the full images rather than only the masked regions. For free-form

extraction, each target photograph is matched separately for each metric to the extracted representative yielding the best score. This matching is performed only for evaluation and does not affect candidate selection.

Main comparisons. For each subject, both masked-input and free-form methods are evaluated on all fine-tuning photographs. Within each run, we average scores over fine-tuning photographs for each subject, then over subjects within each dataset. This gives every subject equal weight regardless of its fine-tuning set size. For A-ESR, we first threshold each photograph's SSCD similarity, then apply the same averaging procedure. Tables 1 and 2 report the mean and sample standard deviation of the dataset-level scores across five independent runs. Each run independently retrains all checkpoints and generates new candidates using different training and sampling seeds. Subject selections, image splits, and mask assignments remain fixed across runs.

Variation across subjects. The run-to-run standard deviation in the main tables measures how much the reported scores vary when training and sampling are repeated. Appendix B additionally reports the mean and sample standard deviation of subject-level scores within each dataset, computed from a single run. This spread describes variation in reconstruction performance across subjects, rather than variability across repeated runs.

Sweeps and held-out evaluation. The mask-coverage and generation-budget sweeps each use a single run, with means and standard deviations computed across subjects. These single-run results need not coincide with the five-run averages in Table 1. Held-out evaluation also uses a single run, with one held-out photograph per subject evaluated under all ten masks in each tier and $N = 20$ candidates generated per masked input. For held-out A-ESR, each subject's score is the fraction of its ten masked reconstructions whose SSCD similarity exceeds the threshold. For each tier, scores are first averaged over masks within each subject, then summarized by the mean and sample standard deviation across subjects.

B Full Experimental Results

B.1 Main Reconstruction Results

We report the full reconstruction results from the first of the five independent runs on all three datasets, including A-ESR@0.7, which is omitted from Table 1. These results supplement Table 1 with variation across subjects within a single run.

Table 3: All metrics are reported as mean ± sample standard deviation across subjects within each dataset, using the first of the five independent runs. These standard deviations describe variation across subjects, whereas those in Table 1 describe variation across runs.

Dataset	Method	AS ↑	A-ESR@0.7 ↑	A-ESR@0.8 ↑	CLIP ↑	LPIPS ↓	SSIM ↑	ArcFace ↑
CelebA-HQ	CFG	42.6±8.2	13.7±12.9	1.3±4.6	86.3±3.4	47.2±5.6	40.1±5.1	69.2±5.6
	FineXtract	53.1±7.9	32.7±17.5	14.3±16.5	89.4±2.7	42.1±5.4	40.0±5.4	77.2±4.3
	Base Inpainting	65.8±5.8	39.0±24.4	6.7±11.2	81.0±3.5	19.7±2.8	68.8±4.0	18.6±5.2
	Personalized Inpainting	88.7±3.2	98.7±4.6	91.3±12.7	96.0±1.0	10.0±1.7	77.1±3.9	82.0±3.8
	PIAR	**91.1±2.6**	**99.0±4.0**	**97.0±7.3**	**96.5±0.8**	**8.5±1.2**	**78.6±3.6**	**87.4±2.3**
DreamBooth	CFG	44.6±10.4	16.9±15.8	3.2±7.3	91.0±3.2	56.7±6.7	32.0±17.4	–
	FineXtract	48.3±10.3	22.9±20.2	7.0±10.8	91.8±2.7	54.6±7.5	30.2±17.5	–
	Base Inpainting	47.3±9.1	4.9±9.4	0.6±3.0	80.0±7.3	30.4±4.7	54.4±15.9	–
	Personalized Inpainting	68.2±12.0	56.6±24.9	26.7±28.8	89.8±7.0	21.6±6.0	59.0±17.9	–
	PIAR	**76.7±7.9**	**78.2±20.9**	**55.8±27.7**	**92.7±5.5**	**17.3±5.1**	**62.8±19.5**	–
CustomConcept101	CFG	40.6±11.7	11.7±17.0	1.7±5.1	90.3±3.6	57.6±6.0	29.0±9.2	–
	FineXtract	45.8±12.0	18.9±17.4	3.3±8.1	91.3±3.3	56.3±6.3	26.9±8.7	–
	Base Inpainting	54.6±7.7	9.4±17.9	0.6±3.0	86.1±6.3	26.3±3.5	58.3±8.5	–
	Personalized Inpainting	72.8±10.7	67.2±32.6	40.6±33.5	93.2±4.6	19.1±4.5	62.4±9.1	–
	PIAR	**78.0±9.5**	**76.1±30.2**	**56.7±34.4**	**94.4±4.0**	**16.6±4.1**	**64.4±9.1**	–

B.2 Checkpoint Specificity Across Datasets

Table 4 extends the checkpoint-specificity comparison in Section 5.3 to all three datasets. Results are reported as mean ± sample standard deviation across five independent runs, following the aggregation used in Table 2.

Table 4: Checkpoint-specificity results across all three datasets. All metrics are reported as mean ± sample standard deviation across five independent runs and are scaled by 100. ArcFace is evaluated only on CelebA-HQ.

Dataset	Method	Target Concept	AS ↑	A-ESR@0.7 ↑	A-ESR@0.8 ↑	CLIP ↑	LPIPS ↓	SSIM ↑	ArcFace ↑
CelebA-HQ	Base Inpainting	–	65.8±0.1	38.9±2.0	7.1±0.5	81.0±0.1	19.6±0.1	69.0±0.2	18.9±0.3
	Personalized Inpainting	Present	88.6±0.1	98.3±0.4	91.9±0.8	95.9±0.1	9.9±0.0	77.3±0.1	82.2±0.1
		Absent	64.7±0.1	35.7±2.7	4.9±1.0	79.9±0.3	17.8±0.1	70.9±0.1	18.6±0.4
	PIAR	Present	90.9±0.2	98.6±0.3	96.6±0.7	96.4±0.1	8.5±0.1	78.7±0.1	87.3±0.2
		Absent	60.9±0.2	22.3±0.4	1.3±0.5	76.9±0.2	18.6±0.1	70.6±0.1	15.0±0.2
DreamBooth	Base Inpainting	–	47.7±0.3	4.8±0.9	0.1±0.2	80.2±0.2	30.1±0.2	54.5±0.1	–
	Personalized Inpainting	Present	68.1±0.2	57.0±0.7	29.2±1.7	90.1±0.2	21.5±0.1	59.2±0.2	–
		Absent	48.0±0.2	5.0±1.2	0.4±0.2	80.8±0.4	29.0±0.3	55.3±0.1	–
	PIAR	Present	76.4±0.3	76.2±1.6	56.4±1.3	92.5±0.3	17.4±0.3	62.8±0.2	–
		Absent	46.5±0.4	3.6±0.5	0.4±0.2	80.4±0.2	28.9±0.1	56.2±0.0	–
CustomConcept101	Base Inpainting	–	54.4±0.2	10.7±1.1	0.8±0.3	86.2±0.3	26.5±0.2	58.3±0.1	–
	Personalized Inpainting	Present	72.5±0.3	64.3±2.3	38.9±2.1	93.1±0.1	19.1±0.1	62.4±0.1	–
		Absent	54.0±0.2	10.1±1.3	1.2±1.0	86.3±0.4	25.9±0.1	59.0±0.1	–
	PIAR	Present	78.0±0.2	75.4±1.6	55.4±1.3	94.4±0.1	16.5±0.1	64.5±0.1	–
		Absent	51.8±0.2	8.1±0.7	1.3±0.6	84.9±0.2	26.6±0.1	59.4±0.1	–

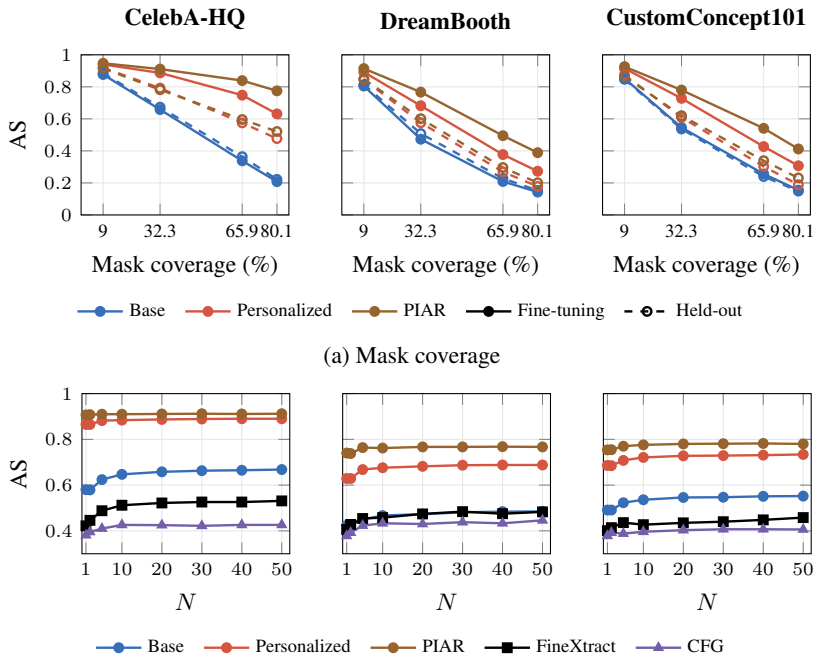
Table 5 reports the corresponding results from the first run, with mean ± sample standard deviation across subjects within each dataset. These results additionally show how checkpoint specificity varies across subjects.

Table 5: Checkpoint-specificity results across all three datasets from the first of the five independent runs. Metrics are scaled by 100; ± denotes sample standard deviation across subjects within each dataset. ArcFace is evaluated only on CelebA-HQ.

Dataset	Method	Target Concept	AS ↑	A-ESR@0.7 ↑	A-ESR@0.8 ↑	CLIP ↑	LPIPS ↓	SSIM ↑	ArcFace ↑
CelebA-HQ	Base Inpainting	–	65.8±5.8	39.0±24.4	6.7±11.2	81.0±3.5	19.7±2.8	68.8±4.0	18.6±5.2
	Personalized Inpainting	Present	88.7±3.2	98.7±4.6	91.3±12.7	96.0±1.0	10.0±1.7	77.1±3.9	82.0±3.8
		Absent	64.9±5.8	37.7±24.7	4.3±8.1	79.5±5.5	17.8±2.4	70.9±3.8	18.1±5.7
	PIAR	Present	91.1±2.6	99.0±4.0	97.0±7.3	96.5±0.8	8.5±1.2	78.6±3.6	87.4±2.3
		Absent	61.1±6.1	22.7±19.9	0.7±3.3	76.8±6.7	18.6±2.5	70.6±3.7	14.9±6.3
DreamBooth	Base Inpainting	–	47.3±9.1	4.9±9.4	0.6±3.0	80.0±7.3	30.4±4.7	54.4±15.9	–
	Personalized Inpainting	Present	68.2±12.0	56.6±24.9	26.7±28.8	89.8±7.0	21.6±6.0	59.0±17.9	–
		Absent	48.4±8.8	4.9±10.7	0.6±3.0	81.4±8.8	28.8±5.3	55.5±16.3	–
	PIAR	Present	76.7±7.9	78.2±20.9	55.8±27.7	92.7±5.5	17.3±5.1	62.8±19.5	–
		Absent	47.1±9.0	3.6±8.5	0.6±3.0	80.5±8.6	28.7±5.0	56.2±16.3	–
CustomConcept101	Base Inpainting	–	54.6±7.7	9.4±17.9	0.6±3.0	86.1±6.3	26.3±3.5	58.3±8.5	–
	Personalized Inpainting	Present	72.8±10.7	67.2±32.6	40.6±33.5	93.2±4.6	19.1±4.5	62.4±9.1	–
		Absent	54.0±7.4	10.6±17.8	0.6±3.0	86.6±5.2	26.0±3.7	58.9±8.8	–
	PIAR	Present	78.0±9.5	76.1±30.2	56.7±34.4	94.4±4.0	16.6±4.1	64.4±9.1	–
		Absent	52.0±7.3	8.9±16.2	2.2±7.2	84.9±5.8	26.6±3.8	59.3±9.0	–

B.3 MASK-COVERAGE SWEEP

Tables 6, 7, and 8 report the full reconstruction results for fine-tuning photographs on CelebA-HQ, DreamBooth, and CustomConcept101 across all four mask tiers, complementing the AS curves in Figure 5(a). Results are obtained from a single run and reported as mean ± sample standard deviation across subjects within each dataset. Across all three datasets, PIAR's AS advantage over Personalized Inpainting increases with mask coverage.



(a) Mask coverage

(b) Generation budget

Figure 7: Effect of mask coverage and generation budget on reconstruction across all three datasets. (a) Solid and dashed lines show fine-tuning and held-out photographs, respectively. PIAR improves reconstruction in both settings, particularly at higher mask coverage. (b) PIAR achieves strong reconstruction with a single candidate and quickly saturates, while CFG and FineXtract benefit more from additional generations. For masked-input methods, N denotes the number of reconstructions per masked input. For CFG and FineXtract, N denotes generations per fine-tuning image, giving NN_0 total generations for a fine-tuning set of N_0 images.

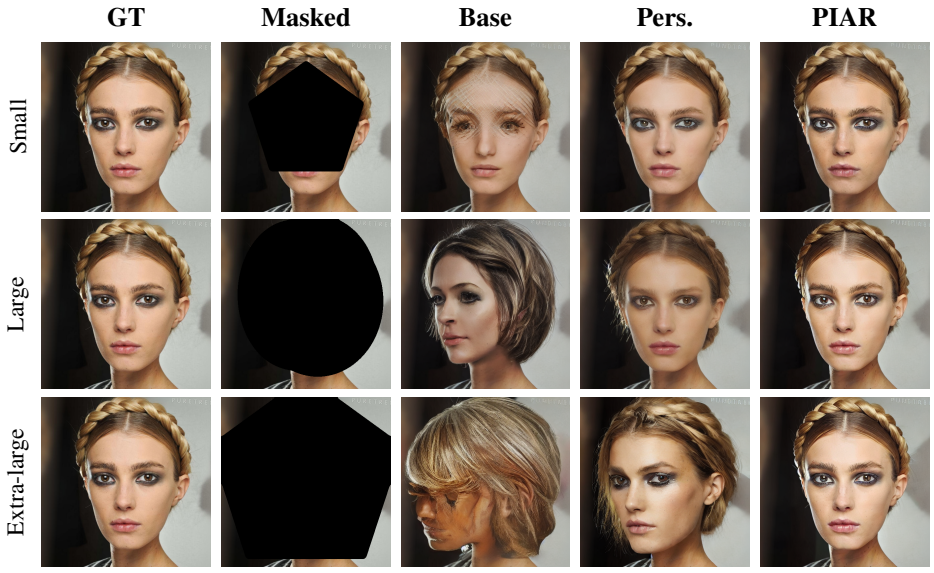


Figure 8: Qualitative effect of mask coverage on reconstruction on CelebA-HQ. Rows show small, large, and extra-large masks, with average coverage of 32.3%, 65.9%, and 80.1%, respectively. GT and Masked denote the ground-truth image and masked input; Base and Pers. denote Base Inpainting and Personalized Inpainting. Personalized Inpainting and PIAR recover the target more faithfully than Base Inpainting, with PIAR preserving more details under large masks.

Table 6: Full reconstruction results for fine-tuning photographs on CelebA-HQ across four mask tiers. Results are reported as mean ± sample standard deviation across 50 subjects within a single run. Metrics are scaled by 100.

Mask tier	Method	AS ↑	A-ESR@0.7 ↑	A-ESR@0.8 ↑	CLIP ↑	LPIPS ↓	SSIM ↑	ArcFace ↑
Partial	Base Inpainting	87.7±2.8	97.3±6.2	87.7±14.6	93.1±1.9	9.0±1.2	80.2±3.3	60.6±6.0
	Personalized Inpainting	94.0±1.8	**100.0±0.0**	98.7±4.6	97.2±0.5	6.7±0.8	82.8±3.0	89.6±2.1
	PIAR	**94.7±1.6**	**100.0±0.0**	**99.7±2.4**	**97.3±0.6**	**6.1±0.7**	**83.4±2.9**	**91.3±1.6**
Small	Base Inpainting	65.8±5.8	39.0±24.4	6.7±11.2	81.0±3.5	19.7±2.8	68.8±4.0	18.6±5.2
	Personalized Inpainting	88.7±3.2	98.7±4.6	91.3±12.7	96.0±1.0	10.0±1.7	77.1±3.9	82.0±3.8
	PIAR	**91.1±2.6**	**99.0±4.0**	**97.0±7.3**	**96.5±0.8**	**8.5±1.2**	**78.6±3.6**	**87.4±2.3**
Large	Base Inpainting	33.8±7.1	0.0±0.0	0.0±0.0	66.6±5.3	43.5±4.7	46.3±4.4	2.4±3.5
	Personalized Inpainting	74.8±6.2	75.3±20.0	38.7±21.4	93.3±2.4	21.0±4.5	63.1±5.1	75.3±6.2
	PIAR	**83.9±3.8**	**94.0±10.5**	**80.0±16.1**	**95.6±1.2**	**15.2±2.9**	**67.2±4.7**	**84.8±3.2**
Extra-large	Base Inpainting	20.8±5.9	0.0±0.0	0.0±0.0	59.1±6.2	56.2±5.3	34.6±4.2	0.7±3.0
	Personalized Inpainting	63.1±9.6	42.3±23.6	9.3±11.7	90.5±4.1	30.7±6.4	53.5±6.1	69.7±8.9
	PIAR	**77.5±6.5**	**81.3±16.4**	**58.0±22.6**	**94.4±2.3**	**21.1±4.2**	**59.9±4.9**	**82.5±5.0**

Table 7: Full reconstruction results for fine-tuning photographs on DreamBooth across four mask tiers. Results are reported as mean ± sample standard deviation across 30 subjects within a single run. Metrics are scaled by 100.

Mask tier	Method	AS ↑	A-ESR@0.7 ↑	A-ESR@0.8 ↑	CLIP ↑	LPIPS ↓	SSIM ↑
Partial	Base Inpainting	80.5±7.6	80.7±22.2	63.2±21.4	94.7±3.2	13.5±3.9	68.0±17.7
	Personalized Inpainting	89.2±6.6	97.2±10.8	91.2±21.1	97.1±1.5	10.3±4.1	70.3±18.2
	PIAR	**91.4±3.3**	**100.0±0.0**	**95.2±15.2**	**97.5±1.3**	**9.0±4.1**	**71.7±18.5**
Small	Base Inpainting	47.3±9.1	4.9±9.4	0.6±3.0	80.0±7.3	30.4±4.7	54.4±15.9
	Personalized Inpainting	68.2±12.0	56.6±24.9	26.7±28.8	89.8±7.0	21.6±6.0	59.0±17.9
	PIAR	**76.7±7.9**	**78.2±20.9**	**55.8±27.7**	**92.7±5.5**	**17.3±5.1**	**62.8±19.5**
Large	Base Inpainting	20.9±7.4	0.0±0.0	0.0±0.0	66.5±7.3	53.9±8.3	34.8±11.1
	Personalized Inpainting	37.7±12.8	5.0±10.9	0.0±0.0	78.3±10.0	42.9±9.5	41.5±15.5
	PIAR	**49.5±15.5**	**23.0±21.0**	**5.9±10.7**	**83.2±10.1**	**37.0±9.1**	**46.2±16.9**
Extra-large	Base Inpainting	14.3±5.7	0.0±0.0	0.0±0.0	62.5±6.2	62.9±8.9	27.1±9.5
	Personalized Inpainting	27.3±11.2	0.6±3.0	0.0±0.0	74.0±10.0	54.1±9.5	32.3±13.1
	PIAR	**38.9±16.4**	**7.7±15.8**	**1.2±4.7**	**79.1±11.2**	**47.4±9.6**	**37.1±15.8**

Table 8: Full reconstruction results for fine-tuning photographs on CustomConcept101 across four mask tiers. Results are reported as mean ± sample standard deviation across 30 subjects within a single run. Metrics are scaled by 100.

Mask tier	Method	AS ↑	A-ESR@0.7 ↑	A-ESR@0.8 ↑	CLIP ↑	LPIPS ↓	SSIM ↑
Partial	Base Inpainting	85.2±3.2	97.2±6.3	77.2±16.1	96.1±2.1	12.6±3.1	70.8±9.4
	Personalized Inpainting	91.5±1.9	99.4±3.0	**98.9±4.2**	97.4±1.0	10.1±2.8	72.7±9.4
	PIAR	**92.6±1.5**	**100.0±0.0**	**98.9±4.2**	**97.5±1.0**	**9.2±2.6**	**73.7±9.5**
Small	Base Inpainting	54.6±7.7	9.4±17.9	0.6±3.0	86.1±6.3	26.3±3.5	58.3±8.5
	Personalized Inpainting	72.8±10.7	67.2±32.6	40.6±33.5	93.2±4.6	19.1±4.5	62.4±9.1
	PIAR	**78.0±9.5**	**76.1±30.2**	**56.7±34.4**	**94.4±4.0**	**16.6±4.1**	**64.4±9.1**
Large	Base Inpainting	25.6±6.2	0.0±0.0	0.0±0.0	71.6±7.5	49.0±5.1	38.7±7.4
	Personalized Inpainting	42.7±12.9	8.9±17.4	2.2±7.2	82.9±8.1	39.4±6.0	44.2±8.7
	PIAR	**54.1±14.5**	**28.9±29.7**	**12.2±16.3**	**86.1±8.1**	**33.9±6.8**	**48.2±9.0**
Extra-large	Base Inpainting	15.5±4.7	0.0±0.0	0.0±0.0	64.2±7.1	58.9±4.6	30.0±6.2
	Personalized Inpainting	30.7±12.7	2.2±5.8	0.6±3.0	77.8±8.9	50.1±6.0	34.9±7.8
	PIAR	**41.2±14.9**	**12.8±22.6**	**3.9±9.5**	**82.0±9.2**	**45.1±6.9**	**39.0±8.8**

B.4 Held-Out Photographs

Tables 9, 10, and 11 report the full reconstruction results for held-out photographs on CelebA-HQ, DreamBooth, and CustomConcept101 across all four mask tiers. Each target depicts the personalized subject but was excluded from the checkpoint's fine-tuning set, following the splits described in Appendix A.1.

Each subject contributes one held-out photograph evaluated under all ten masks in each tier. Results are obtained from a single run. Scores are averaged over masks within each subject and reported as mean ± sample standard deviation across subjects.

Across all three datasets and mask tiers, both Personalized Inpainting and PIAR achieve higher AS than Base Inpainting. PIAR further improves AS over Personalized Inpainting under large and extra-large masks on all three datasets.

Table 9: Full reconstruction results for held-out photographs on CelebA-HQ across four mask tiers. Results are reported as mean ± sample standard deviation across 50 subjects within a single run. Metrics are scaled by 100.

Mask tier	Method	AS ↑	A-ESR@0.7 ↑	A-ESR@0.8 ↑	CLIP ↑	LPIPS ↓	SSIM ↑	ArcFace ↑
Partial	Base Inpainting	88.5±3.5	99.2±3.4	91.8±14.4	93.4±2.0	8.6±1.3	80.3±5.3	61.4±4.6
	Personalized Inpainting	**91.9±3.1**	**100.0±0.0**	**99.0±3.6**	**96.2±1.4**	7.3±1.3	**81.6±5.3**	83.0±5.0
	PIAR	91.5±3.3	99.8±1.4	98.0±8.8	95.9±1.6	**7.2±1.3**	81.5±5.3	**83.6±5.6**
Small	Base Inpainting	67.3±7.1	42.0±29.0	9.0±14.5	82.2±5.3	18.5±2.5	69.1±5.6	19.1±5.4
	Personalized Inpainting	**79.2±7.8**	**85.4±24.1**	**54.8±34.8**	**93.5±3.1**	**13.0±2.5**	**74.0±5.9**	69.4±9.7
	PIAR	78.2±8.9	83.2±27.8	49.0±37.7	93.2±3.2	13.1±3.0	73.7±6.1	**73.0±9.8**
Large	Base Inpainting	36.4±8.6	0.2±1.4	0.0±0.0	68.3±6.7	41.3±4.2	47.3±5.6	2.6±3.3
	Personalized Inpainting	57.5±10.9	16.2±27.2	3.0±11.6	89.7±4.5	27.6±4.7	57.9±7.1	62.0±12.1
	PIAR	**59.6±12.5**	**21.6±32.7**	**8.0±22.6**	**90.5±4.1**	**25.9±5.0**	**58.4±7.0**	**69.1±11.3**
Extra-large	Base Inpainting	22.2±7.1	0.0±0.0	0.0±0.0	59.8±7.1	54.6±4.9	35.5±5.2	1.0±2.9
	Personalized Inpainting	47.7±11.9	4.8±13.7	0.4±2.8	87.4±5.3	37.2±6.7	49.1±7.7	60.1±12.6
	PIAR	**52.1±13.1**	**11.0±25.0**	**4.2±19.0**	**89.1±4.7**	**33.3±6.4**	**50.3±7.5**	**68.8±10.8**

Table 10: Full reconstruction results for held-out photographs on DreamBooth across four mask tiers. Results are reported as mean ± sample standard deviation across 30 subjects within a single run. Metrics are scaled by 100.

Mask tier	Method	AS ↑	A-ESR@0.7 ↑	A-ESR@0.8 ↑	CLIP ↑	LPIPS ↓	SSIM ↑
Partial	Base Inpainting	81.4±9.9	85.3±25.2	66.7±28.9	94.0±3.8	13.0±4.2	69.0±15.9
	Personalized Inpainting	84.6±8.2	91.0±21.9	**78.0±26.6**	95.4±2.7	11.6±4.3	70.1±16.0
	PIAR	**84.9±6.4**	**92.0±15.2**	77.7±27.6	**95.9±2.2**	**11.1±4.4**	**71.0±16.2**
Small	Base Inpainting	50.8±12.5	6.0±9.7	0.3±1.8	81.5±7.8	28.8±5.2	56.0±14.0
	Personalized Inpainting	57.6±13.8	20.3±23.6	3.3±13.7	87.3±7.7	25.1±6.7	58.3±15.0
	PIAR	**60.1±12.0**	**24.7±31.7**	**6.0±19.6**	**89.5±7.1**	**23.3±6.2**	**60.3±15.9**
Large	Base Inpainting	22.8±8.4	0.0±0.0	0.0±0.0	67.7±8.2	53.5±8.0	35.3±9.8
	Personalized Inpainting	26.9±9.7	0.0±0.0	0.0±0.0	73.9±9.5	49.0±7.7	38.3±12.4
	PIAR	**29.6±10.2**	**0.7±3.7**	**0.3±1.8**	**77.0±9.5**	**46.4±7.4**	**41.7±13.7**
Extra-large	Base Inpainting	15.3±6.3	**0.0±0.0**	**0.0±0.0**	62.3±7.5	63.3±8.9	26.3±7.7
	Personalized Inpainting	17.8±7.2	**0.0±0.0**	**0.0±0.0**	68.6±10.5	59.6±8.3	29.3±10.5
	PIAR	**20.1±8.6**	**0.0±0.0**	**0.0±0.0**	**73.0±11.9**	**56.3±7.8**	**33.0±13.1**

Table 11: Full reconstruction results for held-out photographs on CustomConcept101 across four mask tiers. Results are reported as mean ± sample standard deviation across 30 subjects within a single run. Metrics are scaled by 100.

Mask tier	Method	AS ↑	A-ESR@0.7 ↑	A-ESR@0.8 ↑	CLIP ↑	LPIPS ↓	SSIM ↑
Partial	Base Inpainting	84.7±5.0	91.3±13.3	75.7±18.5	95.6±2.9	13.1±3.5	68.1±12.3
	Personalized Inpainting	**87.4±4.5**	**95.3±13.1**	**86.0±15.4**	**96.4±2.6**	11.9±3.3	69.2±12.3
	PIAR	86.5±4.4	94.7±13.1	84.0±18.1	96.3±2.8	**11.7±3.2**	**69.6±12.3**
Small	Base Inpainting	53.7±10.6	11.7±25.7	3.7±11.9	85.3±7.2	27.3±4.6	55.4±11.9
	Personalized Inpainting	60.9±10.1	23.3±35.6	**9.0±19.5**	90.2±7.3	24.1±4.4	57.3±12.2
	PIAR	**62.0±9.7**	**26.0±34.3**	6.7±18.1	**91.3±7.2**	**23.6±4.3**	**58.2±12.2**
Large	Base Inpainting	24.0±5.4	0.0±0.0	**0.0±0.0**	69.2±7.1	50.2±5.5	35.3±8.7
	Personalized Inpainting	30.3±6.6	0.3±1.8	**0.0±0.0**	79.2±8.5	46.0±4.1	38.2±10.4
	PIAR	**33.9±9.7**	**2.3±9.0**	**0.0±0.0**	**82.5±8.6**	**45.0±5.3**	**40.5±11.1**
Extra-large	Base Inpainting	14.9±3.8	**0.0±0.0**	**0.0±0.0**	63.0±7.3	60.1±6.3	26.3±7.5
	Personalized Inpainting	18.9±5.1	**0.0±0.0**	**0.0±0.0**	72.5±9.6	56.7±5.1	28.9±8.6
	PIAR	**23.1±9.3**	**0.0±0.0**	**0.0±0.0**	**77.3±10.0**	**55.9±5.4**	**31.6±9.6**

B.5 Generation Budget

Tables 12, 13, and 14 report AS across all evaluated generation budgets on CelebA-HQ, DreamBooth, and CustomConcept101, complementing Figure 5(b). For masked-input methods, budget N specifies the number of candidates generated per masked input. For CFG and FineXtract, it specifies the number of generations per fine-tuning image, yielding $N \times N_0$ generations per checkpoint, where N_0 is the fine-tuning set size. All masked-input results use the small-mask tier. Results are reported as mean ± sample standard deviation across subjects within each dataset, using a single run.

PIAR achieves strong reconstruction with a single candidate on all three datasets, with only modest gains as the generation budget increases. CFG and FineXtract benefit more from additional generations.

Table 12: Reconstruction performance across generation budgets on CelebA-HQ. AS is reported as mean ± sample standard deviation across 50 subjects within a single run, scaled by 100.

N	CFG	FineXtract	Base Inpainting	Personalized Inpainting	PIAR
1	38.1±7.9	42.2±8.2	58.0±5.3	86.5±4.0	90.7±2.6
2	39.5±7.3	44.6±8.3	57.9±5.2	86.5±4.1	90.8±2.5
5	41.0±7.0	48.8±8.5	62.4±5.7	88.2±3.2	91.0±2.7
10	42.6±7.4	51.2±8.6	64.7±5.6	88.4±3.4	91.0±2.8
20	42.5±8.4	52.2±7.3	65.8±5.8	88.7±3.2	91.1±2.6
30	42.2±8.1	52.6±7.7	66.3±5.8	88.9±3.0	91.2±2.6
40	42.6±8.1	52.6±7.7	66.5±6.1	89.0±3.0	91.1±2.6
50	42.6±8.2	53.1±7.9	66.8±5.9	89.0±3.1	91.2±2.6

Table 13: Reconstruction performance across generation budgets on DreamBooth. AS is reported as mean ± sample standard deviation across 30 subjects within a single run, scaled by 100.

N	CFG	FineXtract	Base Inpainting	Personalized Inpainting	PIAR
1	37.8±9.6	40.6±9.7	42.4±7.9	62.9±12.1	74.0±9.1
2	39.2±9.1	42.8±9.8	42.5±8.0	62.9±12.4	73.8±9.4
5	42.3±10.5	45.4±10.8	45.0±8.7	66.8±11.5	76.4±8.3
10	43.3±10.4	45.8±10.3	46.7±8.7	67.6±11.4	76.2±8.5
20	43.0±11.0	47.4±10.6	47.3±9.1	68.2±12.0	76.7±7.9
30	43.8±11.2	48.4±10.5	48.1±9.1	68.7±11.8	76.7±7.7
40	43.3±9.9	47.5±11.0	48.4±9.0	68.8±12.3	76.8±8.0
50	44.6±10.4	48.3±10.3	48.5±9.1	68.8±12.4	76.7±8.3

Table 14: Reconstruction performance across generation budgets on CustomConcept101. AS is reported as mean ± sample standard deviation across 30 subjects within a single run, scaled by 100.

N	CFG	FineXtract	Base Inpainting	Personalized Inpainting	PIAR
1	37.8±11.0	40.1±11.8	49.1±7.0	68.6±10.8	75.4±10.3
2	39.2±10.8	41.4±12.1	49.1±7.0	68.5±10.5	75.5±10.2
5	38.8±11.1	43.6±11.6	52.3±7.9	70.8±11.2	77.0±10.2
10	39.6±11.1	42.7±11.9	53.6±7.7	72.1±10.4	77.6±9.4
20	40.3±11.1	43.5±11.4	54.6±7.7	72.8±10.7	78.0±9.5
30	40.7±11.2	44.0±11.6	54.7±7.8	72.9±10.5	78.1±9.5
40	40.7±12.3	44.8±12.1	55.1±7.7	73.1±10.4	78.2±9.5
50	40.6±11.7	45.8±12.0	55.2±7.7	73.4±10.1	78.0±9.6

B.6 Additional Qualitative Results

Additional reconstruction examples. Figure 9 presents additional reconstruction examples from CelebA-HQ, DreamBooth, and CustomConcept101, extending the qualitative comparisons in the main text. The examples include faces with varied appearances and subjects from both object datasets.

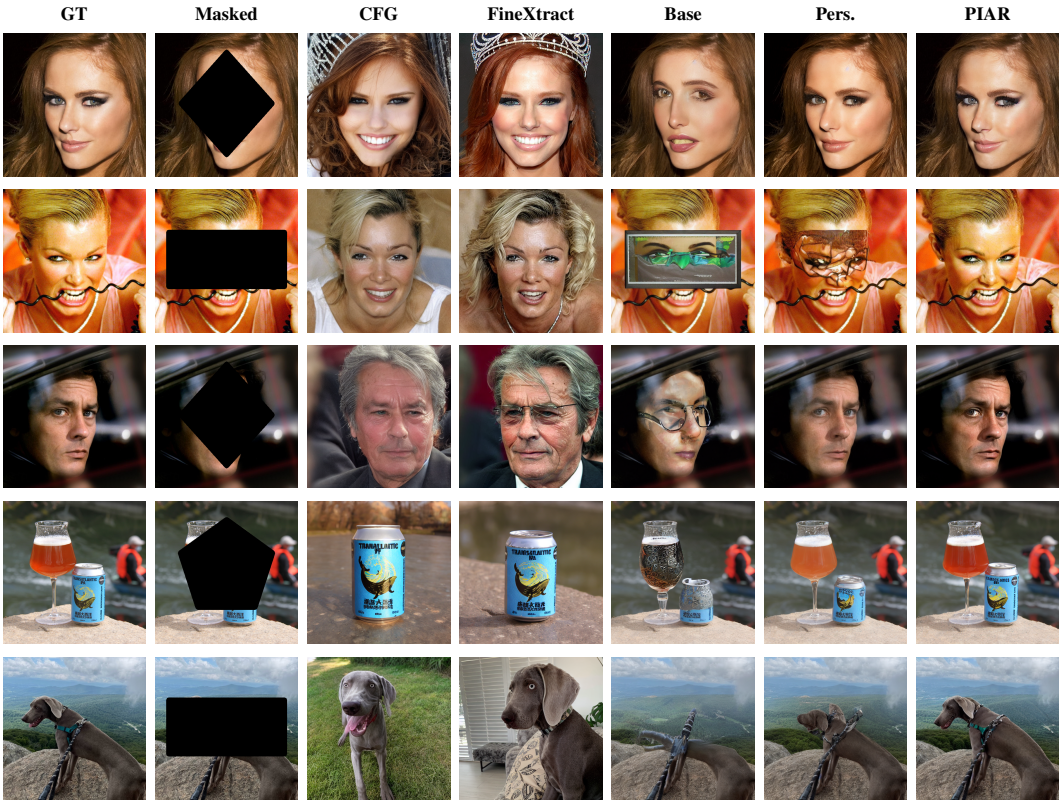


Figure 9: Additional reconstruction examples from CelebA-HQ (rows 1–3), DreamBooth (row 4), and CustomConcept101 (row 5). GT and Masked denote the ground-truth image and masked input; Base and Pers. denote Base Inpainting and Personalized Inpainting, respectively.

Failure cases. Figure 10 shows challenging object examples where PIAR and all compared baselines fail to recover the target's appearance and distinguishing details. In these examples, the masks obscure most of the subject, leaving few visible cues to guide reconstruction.

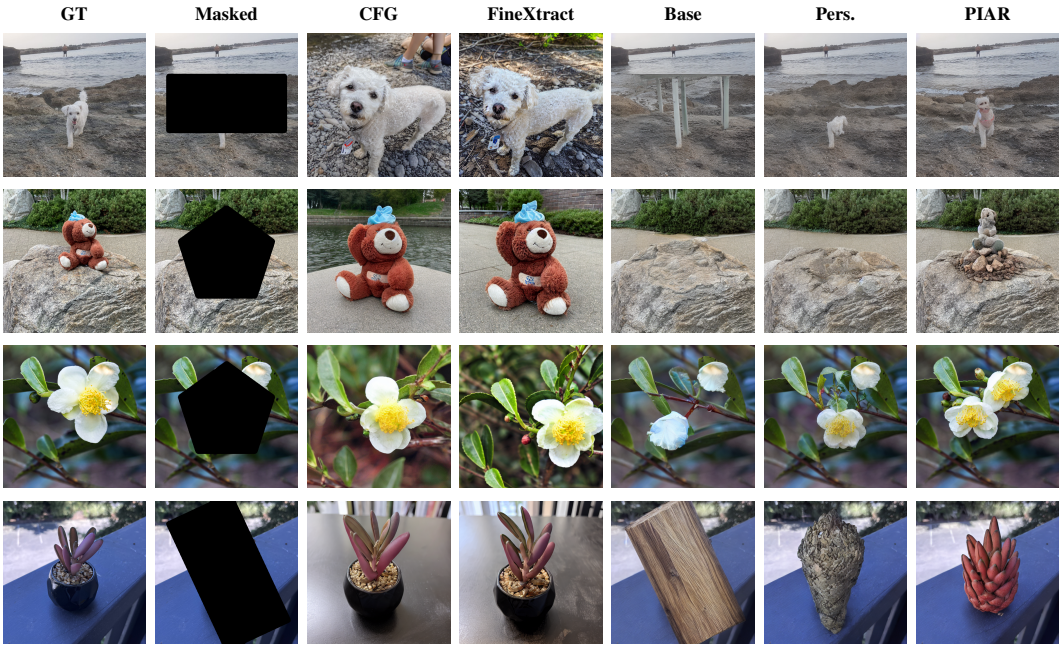


Figure 10: Failure cases on object subjects, where PIAR fails to recover target-specific details despite access to the personalized checkpoint and masked input. GT and Masked denote the ground-truth image and masked input; Base and Pers. denote Base Inpainting and Personalized Inpainting, respectively.

C Additional Studies

C.1 Guidance Scale

We examine sensitivity to w on six randomly selected subjects from each of the three evaluation datasets. Each fine-tuning photograph is paired with a small-tier mask, with $N = 20$ candidates per input and an empty prompt. Table 15 reports the mean and sample standard deviation across subjects within each dataset. AS improves substantially from $w = 1$ to $w = 2$ and remains relatively stable over $w = 2$–5. The highest observed AS occurs at $w = 3$ on CelebA-HQ and CustomConcept101, and at $w = 4$ on DreamBooth. Our default $w = 3$ also matches the guidance weight $w' = 3$ used by FineXtract (Wu et al., 2025).

Table 15: Effect of contrastive guidance weight with an empty prompt. Metrics are scaled by 100 and reported as mean ± sample standard deviation across six subjects within each dataset. $w = 0$ and $w = 1$ correspond to Base Inpainting and Personalized Inpainting, respectively; PIAR uses $w = 3$.

Dataset	w	AS ↑	A-ESR@0.7 ↑	A-ESR@0.8 ↑	CLIP ↑	LPIPS ↓	SSIM ↑
CelebA-HQ	0	66.6±6.6	47.2±32.3	8.3±9.1	79.8±3.6	19.4±3.1	70.6±2.8
	1	89.3±2.5	100.0±0.0	97.2±6.8	96.0±0.8	9.9±1.3	78.7±1.7
	2	92.4±1.7	100.0±0.0	100.0±0.0	96.9±0.7	7.9±0.6	80.6±1.9
	3	92.5±1.5	100.0±0.0	100.0±0.0	96.8±0.5	7.9±0.7	80.4±2.0
	4	92.3±1.4	100.0±0.0	100.0±0.0	96.8±0.5	8.1±0.7	80.2±2.0
	5	91.9±1.4	100.0±0.0	100.0±0.0	96.6±0.4	8.6±0.8	79.7±2.1
DreamBooth	0	46.7±5.6	0.0±0.0	0.0±0.0	76.7±8.3	30.4±4.7	53.9±17.7
	1	67.1±7.5	54.2±21.9	18.6±18.6	86.1±8.3	22.7±5.9	57.9±19.9
	2	72.6±8.1	58.1±22.9	38.3±17.7	88.0±8.3	20.5±6.1	59.6±20.4
	3	72.7±7.2	60.8±17.2	38.3±17.7	89.2±7.7	20.1±5.3	60.2±20.5
	4	74.0±8.6	72.5±19.4	38.3±17.7	89.5±8.0	19.8±5.6	60.2±20.5
	5	73.8±8.9	73.1±28.9	35.0±23.0	88.7±8.2	19.8±5.2	60.3±20.2
CustomConcept101	0	55.6±8.2	16.7±25.8	0.0±0.0	88.4±6.1	28.1±2.5	53.9±10.3
	1	72.2±10.9	61.1±37.5	44.4±34.4	92.6±3.7	21.3±2.6	57.3±10.5
	2	76.8±10.4	72.2±22.8	55.6±34.4	93.7±3.4	19.1±2.0	58.5±10.8
	3	78.6±10.3	77.8±29.2	58.3±34.6	93.6±4.1	18.3±2.0	59.0±10.8
	4	78.3±10.5	77.8±29.2	61.1±36.0	93.9±3.5	18.3±2.3	58.9±10.9
	5	77.6±9.8	77.8±29.2	58.3±34.6	93.5±3.7	18.5±2.2	58.8±10.9

C.2 Effect of the Personalization Prompt

At $w = 3$, we compare an empty prompt with the original training prompt, "a photo of sks1 [class]", using the same subjects and masked inputs for both conditions. As shown in Table 16, providing the training prompt yields slight improvements on object subjects, while substantial reconstruction is already possible with an empty prompt. These results show that access to the training prompt or special token is not necessary for effective reconstruction in this setting.

Table 16: Effect of providing the personalization prompt to PIAR at $w = 3$. Metrics are scaled by 100 and reported as mean ± sample standard deviation across six subjects within each dataset. The training prompt is "a photo of sks1 [class]", where [class] is "person" for CelebA-HQ and the dataset-provided class noun for the object datasets.

Dataset	Prompt	AS ↑	A-ESR@0.7 ↑	A-ESR@0.8 ↑	CLIP ↑	LPIPS ↓	SSIM ↑
CelebA-HQ	Empty (default)	92.5±1.5	100.0±0.0	100.0±0.0	96.8±0.5	7.9±0.7	80.4±2.0
	Training prompt	92.7±1.4	100.0±0.0	100.0±0.0	96.8±0.8	7.7±0.6	80.7±2.2
DreamBooth	Empty (default)	72.7±7.2	60.8±17.2	38.3±17.7	89.2±7.7	20.1±5.3	60.2±20.5
	Training prompt	82.3±2.0	94.4±8.6	63.1±13.5	96.5±2.0	16.8±5.6	61.9±20.7
CustomConcept101	Empty (default)	78.6±10.3	77.8±29.2	58.3±34.6	93.6±4.1	18.3±2.0	59.0±10.8
	Training prompt	85.5±6.8	97.2±6.8	80.6±30.6	96.5±0.8	15.5±2.1	60.6±11.0