# WHEN DIFFUSION MODELS FORGET WHO YOU ARE: IDENTITY PRESERVATION IN FACE INPAINTING UNDER LARGE OCCLUSIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Face inpainting with diffusion models has recently achieved impressive visual quality, yet preserving identity fidelity under significant occlusion and conflicting text guidance remains a major challenge. To address this issue, we present Reference Semantic Inpainting for Face (ReSem-Face), a cascaded diffusion framework that introduces an explicit identity-conditioned semantic prior for multi-reference face inpainting. Our approach distills representative identity features from multiple references to reconstruct missing semantic regions, which then guide the diffusion process through a multi-stream conditioning architecture. This design provides strong semantic constraints when pixels are absent and stabilizes identity reconstruction while remaining compatible with prompt-driven edits. Experiments on CelebAHQ-IDI-5 and VGGFace2 demonstrate that ReSem-Face yields more reliable identity-preserving completion under severe semantic masks and improves text-controlled editing quality compared with representative baselines.

## 1 INTRODUCTION

Face inpainting Ju et al. (2024); Zhuang et al. (2024); Xiao et al. (2025) refers to the task of completing missing or occluded facial regions while preserving the identity of a specific person and maintaining natural consistency with the visible context (e.g., expression, and background). Unlike generic inpainting, the goal is not only to produce a plausible face but to reconstruct the same face: subtle geometry and fine-grained textures that make a person recognizable must be retained even when key identity cues are severely removed by the mask.

Early works typically build on convolutional or GAN-based inpainting pipelinesCai et al. (2019); Li et al. (2017); Goodfellow et al. (2014), sometimes augmented with facial priors such as landmarks, parsing maps, or multi-scale discriminators to enhance realism and structural plausibility. These methods achieve visually coherent completions under mild to moderate occlusions and often produce sharp textures. However, when the missing region is large or covers identity-critical areas (e.g., eyes, nose, and mouth), they tend to hallucinate person-agnostic details, leading to identity drift; moreover, their reliance on local context makes them brittle under expression variations and hard to extend to controllable edits.

It is worth noting that diffusion modelsZhang et al. (2023); Wu et al. (2023); Zhao et al. (2023); Zhu et al. (2024); Wasserman et al. (2025); Xie et al. (2025) have recently advanced image synthesis and editing by progressively denoising from noise to data, offering strong generative priors and impressive fidelity in texture and global consistency. Motivated by this, diffusion-based inpainting frameworks Avrahami et al. (2022; 2023); Zhang et al. (2023); Kim et al. (2025); Huang et al. (2025) have emerged as a promising direction for face restoration, and reference-guided variants further attempt to inject identity information from one or more reference images Luo et al. (2023); Xu et al. (2024)so that the denoising process can reconstruct the correct subject rather than an average face.

Despite the success of reference-guided diffusion and personalized generation techniques, robust identity preservation under heavy occlusion remains challenging. In practice, reference images are often misaligned with the target, as the masked target offers little direct evidence for identity. More importantly, many existing designs inject identity cues implicitly during denoising and expect them to simultaneously determine who the subject is and what content should fill the missing regionRuiz



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.

et al. (2023); Li et al. (2024); Gal et al. (2023); Kumari et al. (2023); under high uncertainty and noise, this coupling can weaken identity signals, and the problem becomes even harder when text-guided editing Zhang et al. (2020b); Nichol et al. (2021); Zhang et al. (2020a); Xie et al. (2023); Chen et al. (2024) is introduced, where semantic instructions may inadvertently override or distort identity-specific details.

Motivated by the aforementioned limitations, we propose Reference Semantic Inpainting for Face (ReSem-Face), a diffusion-based Rombach et al. (2022) personalized face inpainting framework that introduces an explicit identity-conditioned semantic prior to stabilize identity reconstruction under severe occlusion. Specifically, we design an Identity-Aware Semantic Pre-Inpainting module that aggregates cues from multiple references in a clean feature space and predicts semantic tokens describing the missing region's identity-relevant structure and appearance; these tokens are then injected into the diffusion U-Net via dedicated semantic attention, operating alongside text guidance and an introduced Reference Identity Attention pathway to form complementary constraints. By decoupling identity-related semantics from the noisy denoising trajectory, ReSem-Face enables more faithful identity preservation and more reliable controllable editing, especially in challenging large-mask scenarios.

In summary, our contributions are as follows:

- We propose a novel Identity-Aware Semantic Pre-Inpainting Module that predicts identity-aware geometry and texture semantics for missing facial regions, forming a high-level semantic prior for diffusion-based inpainting.
- We integrate this semantic prior into the diffusion U-Net through a Reference Semantic Attention pathway, which works together with text cross-attention and Reference Identity Attention to enable tri-conditioning with text, identity, and semantic guidance.
- Extensive experiments demonstrate that ReSem-Face achieves state-of-the-art identity preservation and strong controllability in both standalone face inpainting and text-guided editing scenarios.

## 2 Related Work

### 2.1 Face Inpainting

The past five years have witnessed rapid progress in face inpainting, driven by the demand to realistically recover occluded facial regions while keeping identity and fine-grained attributes Wang et al. (2021); Xu et al. (2024); Suvorov et al. (2022). Early works focus on identity-guided completion under heterogeneous domains, typically combining staged fitting and refinement to better align the synthesized face with the surrounding context Li et al. (2021); Luo et al. (2023). To this end, Mask-aware transformer (MAT) Li et al. (2022); Vaswani et al. (2017) utilizes transformer-style long-range modeling to fill large holes with coherent global structure at high resolution. Due to the development of deep generative models, diffusion-based priors have been adopted to improve semantic plausibility and output diversity without retraining for specific mask types Lugmayr et al. (2022); Chen et al. (2024); Luo et al. (2023); Rombach et al. (2022). Some works Dong et al. (2022); Luo et al. (2023) propose to explicitly restore structural cues (e.g., sketches and edges) and inject them into texture completion, strengthening geometry–texture consistency. Furthermore, Suvorov et al. (2022) proposes to use Fourier-based global receptive fields to better handle large irregular masks and high-resolution images. However, despite these advances, person-specific fidelity can still drift—subtle identity cues may change when the missing region is large—motivating more identity-aware face inpainting designs Motamed et al. (2023).

### 2.2 Reference-guided Face Inpainting

A widely adopted and highly effective solution to improve identity fidelity in face inpainting is to condition the model on one or multiple reference images of the same subjectZhou et al. (2021); Luo et al. (2023); Motamed et al. (2023); Varanka et al. (2024), which provide identity-critical cues that are absent in heavily masked targets. In early representative efforts, reference-attention–based designs Yu et al. (2022) explicitly align and fuse reference features with the corrupted input, encouraging the completion to resemble the reference identity more faithfully. Following that, dual-control

formulations further disentangle the reference signal into high-level identity and low-level texture Luo et al. (2023), enabling more controllable and higher-quality completion under large-scale missing regions. PVA (Xu et al. 2024) integrates Parallel Visual Attention into a pretrained diffusion inpainting model, injecting reference-image features into the denoising network to achieve identity-preserving and language-controllable face inpainting with lightweight per-identity tuning. Recent works (Yang et al. 2023; Chen et al. 2024a; Motamed et al. 2023) also conduct reference guidance via personalization-by-tuning and diffusion-based exemplar conditioning for stronger reference-driven generation and editing; meanwhile, reference-guided methods have also been extended to directional and diverse face inpainting to produce multiple plausible yet reference-consistent outcomes. Nevertheless, most existing reference-guided methods mainly introduce identity cues as denoising-time conditions, while the semantic content of heavily occluded regions is still inferred implicitly during diffusion. In contrast, our ReSem-Face explicitly predicts identity-aware semantic priors before diffusion denoising, providing additional geometry and texture constraints for large-mask identity reconstruction.

# 3 METHODS

We begin by noting that, in multi-reference identity-preserving face inpainting, a plausible synthesis needs to be generated for a large missing facial region while remaining faithful to the identity of a specific person provided by a set of reference images. This setting becomes even more challenging when the completion is further guided by a text prompt for attribute editing, since identity consistency and prompt compliance can be conflicting under severe occlusions. In this section, we will delve into the details of our proposed ReSem-Face framework, and the pipeline of ReSem-Face is shown in Fig. 1. We first provide a brief review of latent diffusion inpainting, and then present the identity-aware semantic pre-inpainting module, the Reference Identity Attention pathway, and the ReSemAttn-based injection scheme for semantic priors, followed by the training objectives and implementation details.

## 3.1 PRELIMINARIES

Diffusion-based generative models view image synthesis as learning to reverse a gradual noising process. Starting from a clean sample $x_0$, the forward process constructs a sequence $\{x_t\}_{t=1}^{T}$ by repeatedly injecting Gaussian noise with a variance schedule $\{\beta_t\}_{t=1}^{T}$:

$$x_{t+1} = \sqrt{1-\beta_t}\, x_t + \sqrt{\beta_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $0 < \beta_t < 1$. After sufficiently many steps, $x_t$ becomes nearly Gaussian and the original data is largely destroyed. By composing the forward transitions, one can directly relate $x_t$ to $x_0$ in closed form:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon_t, \tag{2}$$

with $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$ denoting the cumulative signal decay. The generative model is then defined by a denoising network $\epsilon_\theta(x_t, t)$ that predicts the noise added at each step. It is trained using a denoising score-matching objective

$$\mathcal{L}_{\mathrm{DSM}} = \mathbb{E}_{x_0, t, \epsilon\sim\mathcal{N}(0,I)}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right], \tag{3}$$

where $t$ is sampled uniformly from $\{1, \ldots, T\}$ and $x_t$ is obtained from $x_0$ via the forward process. At inference time, the model approximately inverts the diffusion chain, starting from Gaussian noise and iteratively denoising it. In practice, we adopt a DDIM-style sampler for efficient generationSong et al. (2020).

Running this procedure directly in pixel space is computationally demanding. Latent Diffusion Models (LDMs)Rombach et al. (2022) address this limitation by applying the same diffusion formulation in a compressed latent representation. Let $E_V(\cdot)$ and $D_V(\cdot)$ denote the encoder and decoder of a Variational Auto-Encoder (VAE) Kingma & Welling (2013), and $z_t = E_V(x_t)$, $x_t = D_V(z_t)$ be the encoding and decoding operations. The diffusion model is then defined over latent variables $z_t$ instead of raw images. Moreover, LDMs are typically conditioned on text through cross-attention layers that attend to language features $y_i = E_T(T_i)$ extracted by a pretrained text encoder (e.g.,

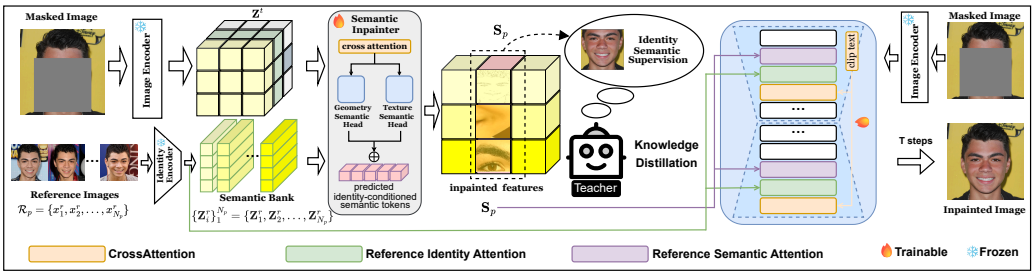



Figure 1: The pipeline consists of three main components: (1) Identity Semantic Bank Construction, which aggregates identity features from multiple reference images; (2) Semantic Pre-Inpainting, where a semantic inpainter predicts semantic tokens for the masked region in the clean feature space; and (3) Tri-Conditioned Diffusion, where the predicted tokens are injected into the denoising backbone via the Reference Semantic Attention (ReSemAttn) pathway, which operates in parallel with text cross-attention and Reference Identity Attention to guide the generation.

CLIPRadford et al. (2021)). For inpainting, the latent model also receives the occluded image and the corresponding binary mask as additional inputs. Concretely, we form a mask-aware latent input

$$\tilde{z}_t = z_t \,\|\, u_{\downarrow}(m) \,\|\, E_V(m \odot x_0), \tag{4}$$

where $m$ is the mask, $u_{\downarrow}(m)$ denotes downsampling $m$ to the latent resolution, $\odot$ is element-wise multiplication, and $\|$ indicates channel-wise concatenation. The inpainting LDM is then trained with a noise-prediction loss analogous to the DDPM objective:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z_0,y,m,t,\epsilon}\big[\|\epsilon - \epsilon_\theta(\tilde{z}_t, y, t)\|_2^2\big], \tag{5}$$

where $z_0 = E_V(x_0)$ and $y$ denotes the text conditioning. Our method builds on top of such a latent diffusion inpainting backbone, which jointly exploits the text prompt, the masked image, and the mask to guide the denoising process.

### 3.2 Identity-Aware Semantic Pre-Inpainting Module

To provide the diffusion backbone with high-level priors describing the latent structure of the occluded facial region, we introduce an Identity-Aware Semantic Pre-Inpainting Module. This module forms an additional semantic pathway complementary to Reference Identity Attention, producing identity-conditioned semantic tokens that encode both geometry and fine-grained appearance. The overall design follows a cascaded Vaswani et al. (2017) architecture reminiscent of CAT-Diffusion but differs fundamentally in its purpose, semantic targets, and integration strategy. The full framework of ReSem-Face is shown in Fig. 1

#### 3.2.1 Multi-Reference Identity Semantic Aggregation

Given a set of reference images $\mathcal{R}_p = \{x_i^r\}_{i=1}^{N_p}$ from the same identity, we first extract per-reference identity-aware tokens using a frozen identity encoder $E_{\text{id}}$, where the FaceNet branch is implemented as a ResNet-50 model trained with ArcFace Deng et al. (2019). To suppress intra-identity variations, we perform cross-image token interaction via multi-head attention over all reference tokens, aggregating them into a unified identity semantic bank:

$$\mathbf{Z}^r = \text{Norm}\left(\frac{1}{N_p}\sum_{i=1}^{N_p} \text{MSA}\big(\mathbf{Z}_i^r, \mathbf{Z}_{1:N_p}^r, \mathbf{Z}_{1:N_p}^r\big)\right), \tag{6}$$

where $\mathbf{Z}_i^r = E_{\text{id}}(x_i^r)$, and $\mathbf{Z}^r$ serves as a consistent identity memory queried by the masked target in subsequent stages.

#### 3.2.2 Identity-conditioned semantic pre-inpainting

Given a masked target image $x^t$ and its binary mask $m$, we encode the visible context with a CLIP-based mask-aware visual encoder $E_{\text{vis}}$ to obtain target tokens $\mathbf{Z}^t$. We then fuse $\mathbf{Z}^t$ with the identity

bank $\mathbf{Z}^r$ through multi-head cross-attention, so that identity cues can be propagated into partially observed or fully missing regions. The fused representation is decoded by a lightweight transformer decoder with two heads: (i) a geometry head that predicts structure-oriented semantics (e.g., component layout and shape consistency), and (ii) a texture head that predicts identity-dependent appearance traits. Their outputs are finally projected into a fixed-length sequence of identity-conditioned semantic tokens $\mathbf{S}_p$ for the occluded region.

#### 3.2.3 IDENTITY-AWARE SEMANTIC SUPERVISION

To ensure $\mathbf{S}_p$ captures both semantic correctness and identity consistency, we supervise the geometry/textural predictions and enforce identity alignment using features extracted from the ground-truth region $x^{\mathrm{gt}}$. In addition, we adopt a frozen teacher model to provide stronger semantic targets and distill high-level knowledge into $\mathbf{S}_p$, which stabilizes semantic prediction under large occlusions and improves prompt alignment:

$$\begin{aligned}\mathcal{L}_{\mathrm{pre}} = \lambda_{\mathrm{id}}\big(1-\cos\big(\phi_{\mathrm{id}}(x^{\mathrm{gt}}),\,\phi_{\mathrm{id}}(\mathbf{S}_p)\big)\big) + \lambda_{\mathrm{geo}}\left\|\phi_{\mathrm{geo}}(x^{\mathrm{gt}})-\mathbf{G}\right\|_1 \\ + \lambda_{\mathrm{tex}}\left\|\phi_{\mathrm{tex}}(x^{\mathrm{gt}})-\mathbf{T}\right\|_1 + \lambda_{\mathrm{t}}\left\|\psi_{\mathrm{t}}(x^{\mathrm{gt}})-\psi_{\mathrm{t}}(\mathbf{S}_p)\right\|_1 .\end{aligned} \tag{7}$$

where $\phi_{\mathrm{id}}, \phi_{\mathrm{geo}}, \phi_{\mathrm{tex}}$ extract identity/geometry/texture descriptors, $\mathbf{G}$ and $\mathbf{T}$ denote the geometry/texture head outputs, and $\psi_{\mathrm{t}}(\cdot)$ denotes teacher features used for distillationNavaneet et al. (2022).

#### 3.2.4 REFERENCE IDENTITY ATTENTION

To complement the semantic prior with direct identity evidence, we introduce a Reference Identity Attention (RIA) pathway for identity-conditioned denoising. Given the aggregated reference identity tokens $\mathbf{H}_p$, RIA lets the intermediate U-Net features selectively retrieve identity-specific cues from the reference set and injects them into the denoising process as an identity guidance branch. Different from the semantic pathway, which predicts explicit geometry and texture priors for the missing region, RIA focuses on preserving subject-level identity consistency during generation. In this way, identity retrieval and semantic prior injection are modeled as two complementary conditions within the tri-conditioning framework.

#### 3.2.5 REFERENCE SEMANTIC ATTENTION FOR DIFFUSION CONDITIONING

We inject the predicted semantic tokens $\mathbf{S}_p$ into the diffusion U-Net via a dedicated Reference Semantic Attention (ReSemAttn) pathway. At each transformer block, ReSemAttn operates in parallel with text cross-attention and Reference Identity Attention (RIA), providing complementary constraints to guide denoising:

$$\begin{aligned}\mathbf{Z}_{\ell+1} = \mathrm{SelfAttn}(\mathbf{Z}_\ell) + \mathrm{CrossAttn}_{\mathrm{text}}(\mathbf{Z}_\ell, \mathbf{Y}) \\ + \mathrm{RIA}(\mathbf{Z}_\ell, \mathbf{H}_p) + \mathrm{ReSemAttn}(\mathbf{Z}_\ell, \mathbf{S}_p),\end{aligned} \tag{8}$$

where $\mathbf{Z}_\ell$ is the hidden state at block $\ell$, $\mathbf{Y}$ is the text condition, and $\mathbf{H}_p$ denotes the reference identity tokens. This tri-pathway conditioning encourages identity-faithful and semantically coherent completion, especially when visible evidence is sparse.

### 3.3 TRAINING

Following the standard latent diffusion inpainting formulation, ReSem-Face optimizes a diffusion-based inpainting objective augmented with reference identity conditioning and identity-aware semantic supervision.

#### 3.3.1 DIFFUSION LOSS

We adopt the standard latent diffusion objective, where the model predicts noise added at timestep $t$. Given masked latent $\tilde{\mathbf{z}}_t$, text prompt $\mathbf{y}$, reference identity tokens $\mathbf{H}_p$, and semantic tokens $\mathbf{S}$, the denoising loss is:

$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}_{\mathbf{z}_0,t,\epsilon}\left[\|\epsilon-\epsilon_\theta(\tilde{\mathbf{z}}_t, t, \mathbf{y}, \mathbf{H}_p, \mathbf{S})\|_2^2\right]. \tag{9}$$

This objective aligns all three conditioning streams (text, identity, semantic) with the diffusion trajectory.

#### 3.3.2 Identity Preservation Loss

To ensure that personalized details are preserved in the generated face, we introduce an *identity-consistency loss* between the reconstructed image $\hat{x}$ and the ground-truth $x^{\mathrm{gt}}$. Using a pretrained identity extractor $\phi_{\mathrm{id}}$, we compute:

$$\mathcal{L}_{\mathrm{id}} = 1 - \cos\big(\phi_{\mathrm{id}}(\hat{x}), \phi_{\mathrm{id}}(x^{\mathrm{gt}})\big)\,. \tag{10}$$

This constraint guides the diffusion model toward identity-faithful synthesis, especially in large-missing-region cases.

#### 3.3.3 Reference Semantic Loss

Our Identity-Aware Semantic Pre-Inpainting Module predicts geometry and texture semantic tokens $(G, T)$ for the occluded region. To align predicted semantics with the ground truth region's structure and appearance, we propose a *Reference Semantic Loss*:

$$\mathcal{L}_{\mathrm{sem}} = \lambda_1 \left\|\phi_{\mathrm{geo}}(x^{\mathrm{gt}}) - G\right\|_1 + \lambda_2 \left\|\phi_{\mathrm{tex}}(x^{\mathrm{gt}}) - T\right\|_1. \tag{11}$$

These terms supervise the semantic branch to produce accurate high-level priors before diffusion denoising.

## 4 Experiments

### 4.1 Experiment Settings

#### 4.1.1 Dataset and Pre-processing

We conduct experiments on CelebAHQ-IDI-5 Xu et al. (2024), a tailored benchmark for multi-reference face inpainting containing 1,963 identities, and VGGFace2 Cao et al. (2018), a large-scale dataset comprising over 3.3 million images across 9,131 identities. Unlike the aligned CelebA-HQ, VGGFace2 serving as a challenging testbed to evaluate our model's generalization capability in unconstrained "in-the-wild" scenarios. Following standard protocols, we utilize 5 reference images per identity and evaluate on unseen identities. To mimic real-world occlusions, we apply semantic masks including lower-face, eye&brow, whole-face, and random regions. The input consists of a masked target image, the corresponding binary mask, and the reference set. All images are kept with original alignment and resized to the diffusion backbone's input resolution. During training, we randomly sample masks to enhance robustness, while evaluation is performed on specific mask categories to assess performance under diverse occlusion semantics.

#### 4.1.2 Implementation Details

ReSem-Face follows a two-stage training scheme. Stage I optimizes the semantic pre-inpainter via $\mathcal{L}_{\mathrm{pre}}$ using Adam with a learning rate of $1 \times 10^{-5}$ for 20K iterations on 4 A40 GPUs. Stage II finetunes the diffusion backbone. We freeze the original U-Net layers and the Identity Encoder, and train only the introduced modules, including ReSemAttn, and Reference Identity Attention, using the joint objective of $\mathcal{L}_{\mathrm{diff}}$, $\mathcal{L}_{\mathrm{id}}$, and $\mathcal{L}_{\mathrm{sem}}$. This stage runs for 200K iterations using AdamW with a learning rate of $1.6 \times 10^{-5}$ and a weight decay of $10^{-2}$, employing classifier-free guidance by dropping conditioning with a probability of 0.1. At inference, we perform a lightweight personalization of 40 steps for each target identity.

#### 4.1.3 Evaluation Metrics

We compare ReSem-Face against eight baselines including LDI Rombach et al. (2022), Custom Diffusion Kumari et al. (2023), Textual Inversion Gal et al. (2023), ReF-LDM Hsiao et al. (2024), PVA Xu et al. (2024), TransRef Liu et al. (2025), OmniGen Xiao et al. (2025), and HiFi-Inpaint Liu et al. (2026). To evaluate identity preservation, we report Identity Similarity computed via CosFace Wang et al. (2018) alongside perceptual metrics such as FID Heusel et al. (2017),




Figure 2: Inpainting results of ReSem-Face and baselines on the test set of CelebAHQ-IDI-5 dataset.

KID Bińkowski et al. (2018); Karras et al. (2020), PSNR, SSIM, and LPIPS Zhang et al. (2018). For text controllability assessed on whole-face masks using 15 edit prompts Xu et al. (2024), we measure CLIPScore Hessel et al. (2021), ImageReward Xu et al. (2023), and Facial Attribute Accuracy(Attr-Acc) Liu et al. (2015). Note that TransRef is excluded from text-based evaluations due to its lack of text conditioning.

### 4.2 Identity-Preserving Face Inpainting

#### 4.2.1 Qualitative Results

Fig. 2 compares ReSem-Face with eight baselines on CelebAHQ-IDI-5. LDI often suffers from severe identity drift, while Custom Diffusion and PVA improve identity consistency but frequently yield over-smoothed textures or structural inconsistencies under large occlusions. Although OmniGen and TransRef enhance reference alignment, they still exhibit subtle identity mismatches or boundary artifacts. In contrast, ReSem-Face generates the most faithful completions across all mask types. By leveraging the identity-conditioned semantic prior, our method accurately reconstructs identity-defining structures and sharp textures, significantly outperforming baselines in maintaining identity coherence when visual evidence is sparse.

#### 4.2.2 Quantitative Results

Tab. 1 summarizes the quantitative comparisons on CelebAHQ-IDI-5. First, reference-guided methods substantially outperform generic inpainting and lightweight personalization in identity similarity. Second, while TransRef achieves high PSNR and SSIM due to its pixel-aligned prompting, ReSem-Face attains the best FID and LPIPS. This indicates that our method generates more perceptually natural textures compared to pixel-wise optimization. Most importantly, ReSem-Face achieves the highest ID score of 0.766, surpassing the strongest baseline by a clear margin. These results confirm that the proposed semantic prior effectively enhances both identity fidelity and perceptual realism under large occlusions.

#### 4.2.3 Cross-Dataset Evaluation on VGGFace2

To evaluate generalization capabilities on in-the-wild data, we conduct experiments on the VGGFace2 test set Cao et al. (2018), with qualitative comparisons shown in Fig. 3. We utilize the official test split with faces aligned and resized to $512 \times 512$, noting that the resulting ground-truth images often exhibit blurriness which affects reference-based metrics. Quantitative results in Tab. 2 show that ReSem-Face leads in identity preservation with an ID score of 0.672. Although all methods suffer from domain gaps and resolution mismatches compared to CelebA-HQ, ReSem-Face

Table 1: Quantitative results on the test set of CelebAHQ-IDI-5. We quantify the per-identity tuning overhead in the "FT.Time" column using a single RTX A40, where "ID" denotes the identity similarity score. Fine-tuning is limited to 40 steps for all applicable models. KID is scaled by $10^{-3}$. The best and second-best results are highlighted in **bold** and underline, respectively.

| Method | Venue | FT.Time | ID ↑ | FID ↓ | KID ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|---|---|---|---|---|
| LDI Rombach et al. (2022) | CVPR'22 | – | 0.326 | 8.79 | **2.92** | 24.52 | 0.815 | 0.138 |
| CD Kumari et al. (2023) | CVPR'23 | ∼3min | 0.688 | 15.86 | 7.87 | 26.85 | 0.862 | 0.132 |
| TI Gal et al. (2023) | ICLR'23 | ∼6min | 0.615 | 17.50 | 9.12 | 25.40 | 0.830 | 0.138 |
| ReF-LDM Hsiao et al. (2024) | NeurIPS'24 | – | 0.725 | 8.35 | 4.55 | 27.50 | 0.885 | 0.122 |
| PVA Xu et al. (2024) | WACV'24 | ∼1min | 0.736 | 8.16 | 4.27 | 27.92 | 0.895 | 0.118 |
| TransRef Liu et al. (2025) | Neurocomputing'25 | – | 0.713 | 8.56 | 4.79 | **28.45** | **0.910** | 0.128 |
| OmniGen Xiao et al. (2025) | CVPR'25 | – | 0.704 | 9.45 | 4.82 | 26.14 | 0.848 | 0.126 |
| HiFi-Inpaint Liu et al. (2026) | CVPR'26 | – | 0.722 | 9.79 | 4.91 | 26.34 | 0.896 | 0.125 |
| ReSem-Face (Ours) | – | ∼1min | **0.766** | **7.90** | 3.75 | 28.31 | 0.904 | **0.116** |

Table 2: Cross-dataset evaluation on the test set of VGGFace2. The metric definitions and fine-tuning settings are identical to those in Tab. 1.

| Method | Venue | FT.Time | ID ↑ | FID ↓ | KID ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|---|---|---|---|---|
| LDI Rombach et al. (2022) | CVPR'22 | – | 0.312 | 17.50 | **6.85** | 21.45 | 0.715 | 0.172 |
| CD Kumari et al. (2023) | CVPR'23 | ∼3min | 0.585 | 21.30 | 12.40 | 22.10 | 0.742 | 0.165 |
| TI Gal et al. (2023) | ICLR'23 | ∼6min | 0.525 | 22.80 | 13.50 | 21.50 | 0.725 | 0.175 |
| ReF-LDM Hsiao et al. (2024) | NeurIPS'24 | – | 0.645 | 17.90 | 8.85 | 23.80 | 0.775 | 0.148 |
| PVA Xu et al. (2024) | WACV'24 | ∼1min | 0.638 | 18.10 | 8.95 | 24.15 | 0.782 | 0.145 |
| TransRef Liu et al. (2025) | Neurocomputing'25 | – | 0.612 | 18.85 | 9.55 | 24.12 | **0.795** | 0.152 |
| OmniGen Xiao et al. (2025) | CVPR'25 | – | 0.605 | 19.42 | 10.12 | 23.05 | 0.768 | 0.158 |
| HiFi-Inpaint Liu et al. (2026) | CVPR'26 | – | 0.612 | 17.82 | 8.52 | 23.94 | 0.788 | 0.151 |
| ReSem-Face (Ours) | – | ∼1min | **0.672** | **16.85** | 7.62 | **24.56** | 0.792 | **0.142** |

consistently achieves the lowest FID and LPIPS. This confirms that our framework generalizes well to unseen identities and remains robust even when reference images are of lower quality.

### 4.3 Ablation Study

We validate the effectiveness of the proposed semantic prior and its injection strategy through component-wise ablation (Tab. 3) and hyperparameter analysis (Fig. 4).

#### 4.3.1 Component Analysis

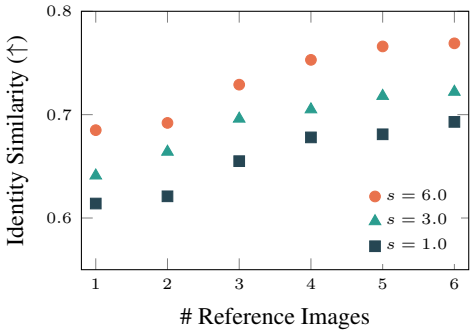
We compare three variants: (1) no semantic prior, (2) semantic prior without injection using auxiliary supervision only, and (3) the full model with ReSemAttn injection. Results in Tab. 3 show that removing the semantic prior entirely leads to a significant performance drop. Furthermore, Var. 2 yields only marginal gains, verifying that explicit injection via ReSemAttn is crucial for stabilizing identity reconstruction. Notably, our full model (Var. 3) ensures stable semantic injection and structural consistency without compromising the fidelity of identity texture, leading to the superior performance observed in ReSem-Face.

Table 3: Component ablation analysis on CelebAHQ-IDI-5. We evaluate the impact of the semantic prior and its injection strategy on both identity-preserving inpainting and text-controlled editing.

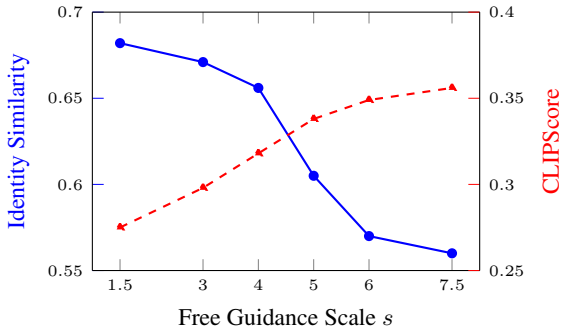
| Components | | | Identity-Preserving | | | Text-Controlled | |
|---|---|---|---|---|---|---|---|
| # | SemPrior | ReSemAttn | ID ↑ | FID ↓ | KID $\times 10^{-3}$ ↓ | ID ↑ | CLIPScore ↑ |
| 1 | | | 0.702 | 9.61 | 5.12 | 0.592 | 0.311 |
| 2 | ✓ | | 0.747 | 9.62 | 5.11 | 0.613 | 0.309 |
| 3 | ✓ | ✓ | **0.766** | **7.90** | **3.75** | **0.656** | **0.318** |




Figure 3: Inpainting results of ReSem-Face and baselines on the test set of VGGFace2 dataset.




(a) Impact of Reference Number




(b) ID vs. CLIP Trade-off

Figure 4: Ablation studies on the number of reference images and free guidance scale. (a) Increasing the number of reference images steadily improves identity consistency across different free guidance scales. (b) Increasing guidance scale improves text alignment (CLIP) but leads to identity degradation (ID); we choose $s = 4.0$ as the optimal trade-off point.

### 4.3.2 Hyperparameter Analysis

We further investigate the impact of key hyperparameters in Fig. 4. First, regarding the number of reference images (Fig. 4(a)), we observe that increasing the references from 1 to 6 consistently boosts identity similarity. Notably, using multi-reference guidance maintains high identity fidelity even under strong guidance scales. Second, regarding the conflict between text and identity (Fig. 4(b)), we analyze the trade-off under varying guidance scales ($s$). While higher scales improve CLIPScore, they often degrade identity preservation. Our analysis identifies $s = 4.0$ as the optimal sweet spot, achieving precise attribute editing with minimal loss in identity fidelity.

## 5 Conclusion

In this paper, we presented ReSem-Face, a semantic-enhanced diffusion framework for identity-preserving and text-controllable face inpainting under large occlusions. By combining identity-aware semantic pre-inpainting, Reference Identity Attention, and Reference Semantic Attention, ReSem-Face provides complementary identity and semantic constraints during denoising. Experiments on CelebAHQ-IDI-5 and VGGFace2 demonstrate its superiority in identity fidelity, visual realism, and text controllability over representative baselines.

### AI USE STATEMENT

In this work, we used generative AI tools for language polishing and grammar checking of the manuscript. We have not used generative AI tools for idea generation, code writing, or data analysis. Additionally, we used generative AI tools for improving the clarity of certain sentences. We have reviewed all AI-assisted work. Specifically, all text modifications suggested by the AI were carefully checked and edited by the authors to ensure accuracy and alignment with our original intent. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

## REFERENCES

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18208–18218, 2022.

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42(4):1–11, 2023.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Jiancheng Cai, Hu Han, Shiguang Shan, and Xilin Chen. Fcsr-gan: Joint face completion and super-resolution via multi-task learning. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):109–121, 2019.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

Yifu Chen, Jingwen Chen, Yingwei Pan, Yehao Li, Ting Yao, Zhineng Chen, and Tao Mei. Improving text-guided object inpainting with semantic pre-inpainting. In *European conference on computer vision*, pp. 110–126. Springer, 2024.

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.

Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11358–11368, 2022.

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 *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NAQvF08TcyG.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Chi-Wei Hsiao, Yu-Lun Liu, Cheng-Kun Yang, Sheng-Po Kuo, Kevin Jou, and Chia-Ping Chen. Ref-ldm: A latent diffusion model for reference-based face image restoration. *Advances in Neural Information Processing Systems*, 37:74840–74867, 2024.

Jun Huang, Ting Liu, Yihang Wu, Xiaochao Qu, Luoqi Liu, and Xiaolin Hu. Mtadiffusion: Mask text alignment diffusion model for object inpainting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18325–18334, 2025.

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*, pp. 150–168. Springer, 2024.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

Sora Kim, Sungho Suh, and Minsik Lee. Rad: Region-aware diffusion models for image inpainting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2439–2448, 2025.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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.

Jia Li, Zhaoyang Li, Jie Cao, Xingguang Song, and Ran He. Faceinpainter: High fidelity face adaptation to heterogeneous domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5089–5098, 2021.

Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10758–10768, 2022.

Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3911–3919, 2017.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8640–8650, 2024.

Taorong Liu, Liang Liao, Delin Chen, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shin'ichi Satoh. Transref: Multi-scale reference embedding transformer for reference-guided image inpainting. *Neurocomputing*, 632:129749, 2025.

Yichen Liu, Donghao Zhou, Jie Wang, Xin Gao, Guisheng Liu, Jiatong Li, Quanwei Zhang, Qiang Lyu, Lanqing Guo, Shilei Wen, et al. Hifi-inpaint: Towards high-fidelity reference-based inpainting for generating detail-preserving human-product images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1994–2004, 2026.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 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.

Wuyang Luo, Su Yang, and Weishan Zhang. Reference-guided large-scale face inpainting with identity and texture control. *IEEE Transactions on Circuits and Systems for Video Technology*, 33 (10):5498–5509, 2023.

Saman Motamed, Jianjin Xu, Chen Henry Wu, Christian Häne, Jean-Charles Bazin, and Fernando De la Torre. Patmat: Person aware tuning of mask-aware transformer for face inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22778–22787, 2023.

KL Navaneet, Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Simreg: Regression as a simple yet effective tool for self-supervised knowledge distillation. *arXiv preprint arXiv:2201.05131*, 2022.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

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.

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*, pp. 22500–22510, 2023.

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

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2149–2159, 2022.

Tuomas Varanka, Tapani Toivonen, Soumya Tripathy, Guoying Zhao, and Erman Acar. Pfstorer: Personalized face restoration and super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2372–2381, 2024.

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

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5265–5274, 2018.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9168–9178, 2021.

Navve Wasserman, Noam Rotstein, Roy Ganz, and Ron Kimmel. Paint by inpaint: Learning to add image objects by removing them first. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18313–18324, 2025.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7623–7633, 2023.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Liangbin Xie, Daniil Pakhomov, Zhonghao Wang, Zongze Wu, Ziyan Chen, Yuqian Zhou, Haitian Zheng, Zhifei Zhang, Zhe Lin, Jiantao Zhou, et al. Turbofill: Adapting few-step text-to-image model for fast image inpainting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7613–7622, 2025.

Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22428–22437, 2023.

Jianjin Xu, Saman Motamed, Praneetha Vaddamanu, Chen Henry Wu, Christian Haene, Jean-Charles Bazin, and Fernando De la Torre. Personalized face inpainting with diffusion models by parallel visual attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5432–5442, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Jiazuo Yu, Kai Li, and Jinjia Peng. Reference-guided face inpainting with reference attention network. *Neural Computing and Applications*, 34(12):9717–9731, 2022.

Lisai Zhang, Qingcai Chen, Baotian Hu, and Shuoran Jiang. Text-guided neural image inpainting. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 1302–1310, 2020a.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

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.

Zijian Zhang, Zhou Zhao, Zhu Zhang, Baoxing Huai, and Jing Yuan. Text-guided image inpainting. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 4079–4087, 2020b.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:11127–11150, 2023.

Yuqian Zhou, Connelly Barnes, Eli Shechtman, and Sohrab Amirghodsi. Transfill: Reference-guided image inpainting by merging multiple color and spatial transformations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2266–2276, 2021.

Rui Zhu, Yingwei Pan, Yehao Li, Ting Yao, Zhenglong Sun, Tao Mei, and Chang Wen Chen. Sd-dit: Unleashing the power of self-supervised discrimination in diffusion transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8435–8445, 2024.

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 Appendix

## A.1 Text Controllability

### A.1.1 Qualitative Results

Fig. 5 presents qualitative comparisons for text-controlled inpainting. Custom Diffusion responds to prompts but compromises identity under strong edits. PVA maintains stability yet over-smoothes complex attributes. OmniGen captures global structure but struggles to balance fine-grained attributes with identity constraints. In contrast, ReSem-Face delivers the most compelling results across prompts, faithfully executing target attributes while preserving identity-defining structures and natural textures, demonstrating that our semantic prior effectively bridges language control and identity preservation.

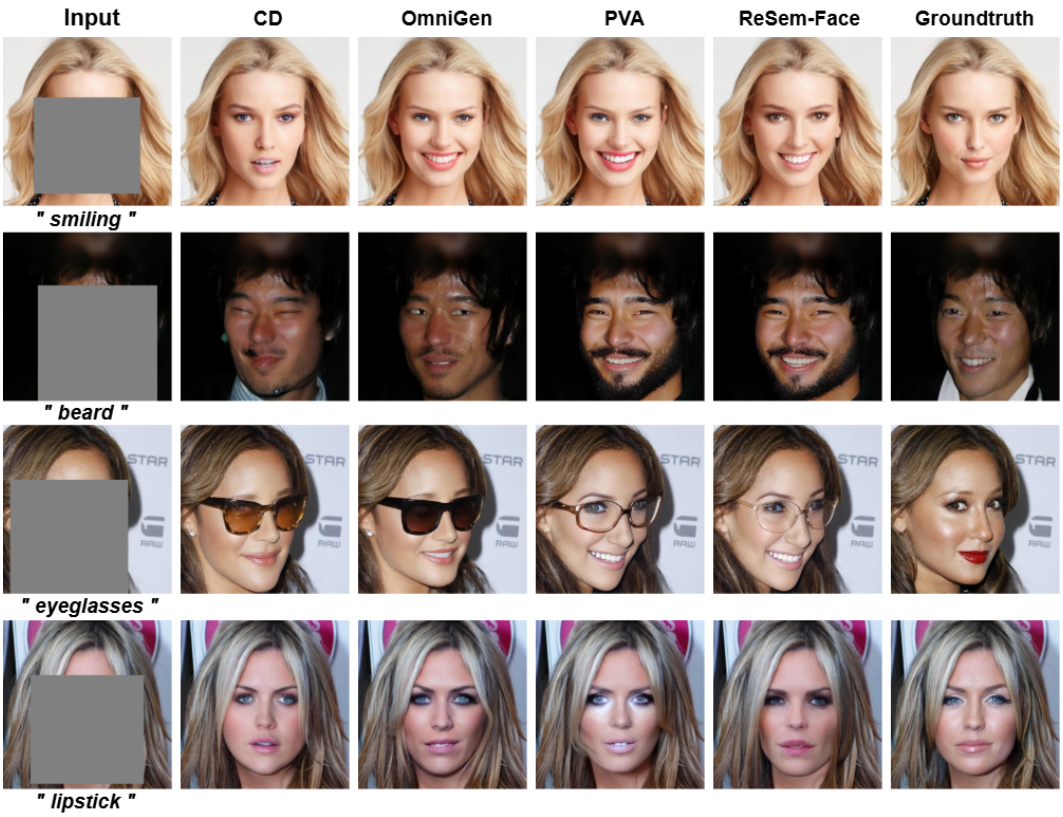



Figure 5: Qualitative comparisons of identity-preserving text-controlled inpainting. Prompts for editing are shown at the bottom of each row. The column tabs, "CD, OmniGen, PVA" denote Custom Diffusion Kumari et al. (2023), OmniGen Xiao et al. (2025), PVA Xu et al. (2024), respectively.

### A.1.2 Quantitative results

Tab. 4 reports quantitative results, including CLIPScore, ImageReward, and Attr-Acc. Baseline methods often sacrifice identity for prompt alignment. In contrast, ReSem-Face achieves the best overall trade-off, obtaining the highest ID similarity, CLIPScore, and Attr-Acc while maintaining competitive ImageReward. These results confirm that our semantic prior enables precise editing without compromising identity consistency.

### A.1.3 User Study

Next, we carry out a user study to examine whether the inpainted images conform to human preferences. In the experiments, we randomly sample 1K images from the test set of CelebAHQ-IDI-5. We invite 10 evaluators (5 males and 5 females) with diverse education backgrounds: art design (4), psychology (2), computer science (2), and business (2). We present the evaluators with the

Table 4: Quantitative results on CelebAHQ-IDI-5 for text-controlled face inpainting. We additionally report ImageReward to evaluate human preference and Attr-Acc for attribute classification accuracy. The best and second-best results are highlighted in **bold** and underline, respectively.

| Method | ID ↑ | CLIPScore ↑ | ImageReward ↑ | Attr-Acc (%) ↑ |
|---|---|---|---|---|
| CD | 0.617 | 0.256 | 0.52 | 81.5 |
| OmniGen | 0.602 | 0.289 | **0.68** | 89.2 |
| PVA | <u>0.632</u> | <u>0.309</u> | 0.61 | <u>91.8</u> |
| ReSem-Face | **0.656** | **0.318** | <u>0.65</u> | **92.5** |

Table 5: User study results on 1K randomly sampled test images from CelebAHQ-IDI-5. We report the average user preference score (1-5), where higher is better.

| Method | LDI | CD | OmniGen | PVA | ReSem-Face |
|---|---|---|---|---|---|
| Identity Fidelity | 1.45 | 2.82 | 3.15 | 3.72 | 4.35 |
| Text Alignment | 2.15 | 2.50 | 3.88 | 3.25 | 4.18 |
| Visual Realism | 2.90 | 3.20 | 3.65 | 3.58 | 4.40 |

masked inputs, reference images, text prompts, and the inpainted results from different methods in a randomized order. We ask them to assign scores (1∼5) from three aspects: 1) Identity Fidelity: whether the completed face maintains the identity characteristics of the reference images; 2) Text Alignment: whether the generated content aligns with the semantic attributes described in the text prompt; 3) Visual Realism: whether the inpainted region is visually coherent with the surrounding context and structurally natural. Table 5 summarizes the averaged results of different approaches. As indicated by the results, ReSem-Face leads the competition by a clear margin against the other baselines in terms of both identity preservation and text-driven editability, aligning best with human perception.