FlowID: Enhancing Forensic Identification with Latent Flow-Matching Models

Jules Ripoll[1] , **David Bertoin**[1] , **Alasdair Newson**[2] , **Charles Dossal**[1] and **Jose Pablo Baraybar**[3]
[1]INSA Toulouse
[2]Sorbonne Université
[3]International Committee of the Red Cross
{jripoll, bertoin, dossal}@insa-toulouse.fr, anewson@isir.upmc.fr, jpbaraybar@icrc.org

Abstract

Every day, many people die under violent circumstances, whether from crimes, war, migration, or climate disasters. Medico-legal and law enforcement institutions document many portraits of the deceased for evidence, but cannot immediately carry out identification on them. While traditional image editing tools can process these photos for public release, the workflow is lengthy and produces suboptimal results. In this work, we leverage advances in image generation models, which can now produce photorealistic human portraits, to introduce FlowID, an identity-preserving facial reconstruction method. Our approach combines single-image fine-tuning, which adapts the generative model to out-of-distribution injured faces, with attention-based masking that localizes edits to damaged regions while preserving identity-critical features. Together, these components enable the removal of artifacts from violent death while retaining sufficient identity information to support identification. To evaluate our method, we introduce InjuredFaces, a novel benchmark for identity-preserving facial reconstruction under severe facial damage. Beyond serving as an evaluation tool for this work, InjuredFaces provides a standardized resource for the community to study and compare methods addressing facial reconstruction in extreme conditions. Experimental results show that FlowID outperforms state-of-the-art open-source methods while maintaining low memory requirements, making it suitable for local deployment without compromising data privacy.

Content warning: This paper contains images of facial injuries that readers may find unpleasant.

1 Introduction

Visual recognition by next-of-kin remains the first and most critical step in identifying human corpses for medico-legal institutions worldwide. In practice, photographs of the deceased are shown to potential relatives, who assess whether they recognize the individual. In cases of violent deaths, said photographs can be extremely graphic and cause severe psychological distress to the person seeing the picture, and may be unnecessary if the recognition ultimately fails. To circumvent this issue, and enable broader dissemination, traditional image editing tools are used to remove visible death-related artifacts. Unfortunately, this method is labor-intensive and only produces satisfying results for minor injuries, failing for severe or structural facial trauma. This limitation is particularly problematic, given the urgent need for efficiency gains in medico-legal workflows. In Mexico alone, more than 72,000 bodies remain unidentified alongside 130,000 reported missing persons.

Figure 1: Samples from InjuredFaces, our facial reconstruction benchmark, edited with FlowID. Top: original images. Bottom: edited results.

Given the scale of data needed to be processed for identification, semi-automatic tools that accelerate and make the identification process easier would provide substantial societal benefit. This motivates the following question : *"Can generative model-based image editing be effectively applied to facial reconstruction to facilitate identification by next-of-kin?"* This question was inspired by efforts by the International Committee of the Red Cross (ICRC) to address the migration crisis across the Mediterranean. Initiatives like "Trace the Face"[1] — a virtual photo gallery enabling people to search for loved ones lost during migration—have reconnected 309 families out of 7,677 published photographs, but remain primarily designed for the living. The challenge of identifying the deceased is far more complex: bodies are often recovered without identification documents or any information concerning their identity

[1]https://tracetheface.familylinks.icrc.org/?lang=en

or whereabouts, and photographs may need to be shared with communities of origin or the diaspora to stimulate recognition. This problem extends beyond migration to other forensic settings where facial trauma impedes visual recognition.

Addressing this challenge has only recently become feasible due to rapid advances in image generation. Modern generative models achieve high visual fidelity and fine-grained semantic control, making them suitable candidates for controlled facial reconstruction. Recent progress is largely driven by diffusion [Sohl-Dickstein *et al.*, 2015; Ho *et al.*, 2020; Song *et al.*, 2021b] and flow-matching models [Liu *et al.*, 2022; Lipman *et al.*, 2023], enhanced by transformer-based architectures [Peebles and Xie, 2023; Esser *et al.*, 2024; Labs, 2024]. Despite these advances, deploying generative image editing in a forensic context poses substantial challenges. First, reconstructed faces must remain as faithful as possible to the original identity except for the death-related artifacts that are intentionally removed. This tension between preserving fidelity and enabling meaningful edits is commonly referred to as the editability-reconstruction trade-off and lies at the core of image editing. Second, processing of such sensitive data must be performed locally, using relatively common hardware, to satisfy privacy and operational constraints. This requirement is non-trivial as editing performance typically scales with model size and computational cost.

In this work, we address both of these challenges and demonstrate that generative image editing can be integrated into forensic identification procedures, to support recognition by next-of-kin. We introduce InjuredFaces, the first benchmark for evaluating generative facial reconstruction under severe damage, and propose FlowID, an algorithm that outperforms other open-source methods while remaining lightweight enough for deployment on consumer hardware. The deployment of our algorithm has started in various locations in South America, enabling the identification of several persons.

2 Related Work

Training-free image editing. Training-free methods mostly rely on perturbing the source image via a forward noising process to project it to a more editable state, and then applying a reverse denoising process to steer the image towards a target edit. SDEdit [Meng *et al.*, 2022] was the first to demonstrate this idea to produce edits where coarse guiding sketches were transformed into realistic samples. However, large edits require sending the source image to strong noise levels to be faithfully applied, which significantly diminishes the fidelity to the input. A major step forward came with the introduction of deterministic solvers, such as in [Song *et al.*, 2021b] and DDIM [Song *et al.*, 2021a]. Beyond accelerating sampling, these solvers allow for image inversion, that is, mapping the image back to its latent representation by reversing the denoising process. When the image is well aligned with the model distribution, inversion yields a likely generation trajectory that can be reused for editing, and this principle has since become the foundation of many training-free approaches. Building on this, several work refine inversion for editing. DiffEdit [Couairon *et al.*, 2022] combines automatic mask generation with DDIM inversion, ensuring perfect reconstruction in unedited regions. Other methods exploit reference cross-attention maps to better control the layout during editing : [Parmar *et al.*, 2023] aligns attention maps through gradient steps, while Prompt-to-Prompt [Hertz *et al.*, 2022] directly substitutes maps from the reference prompt. Null-Text Inversion [Mokady *et al.*, 2023] extends these techniques to real images by optimizing null embeddings during DDIM inversion. More recent approaches further improve fidelity: DDPM inversion [Huberman-Spiegelglas *et al.*, 2024] achieves exact reconstruction by storing noise maps, while LEdits++ [Brack *et al.*, 2024] restricts edits using masks derived from cross-attention. TightInversion [Kadosh *et al.*, 2025] uses network adapters to condition inversion with an image to improve reconstruction quality while retaining editability. RF-Solver [Wang *et al.*, 2024] increases inversion accuracy by employing higher-order ODE solvers, albeit at higher computational cost. Despite their effectiveness, training-free methods often require careful hyperparameter tuning and remain sensitive to inversion quality, which can limit their robustness and reliability in practice. FlowID addresses the issue of inversion quality with a light single-image fine-tuning, achieving a better balance between editability and preservation of the original image, while retaining a relatively low hyperparameter count.

Training-based image editing. This framework takes a different approach by directly learning the parameters of the model dedicated to image editing, which generally translate to better editing performance. ICEdit [Zhang *et al.*, 2025] leverages the inherent capability of modern text-to-image DiTs [Esser *et al.*, 2024] to generate coherent panels and further improves the editing performance by learning LoRA [Hu *et al.*, 2022] parameters to better follow editing instructions. Another research direction focuses on instruction-based image editing. InstructPix2Pix [Brooks *et al.*, 2023], EmuEdit [Sheynin *et al.*, 2024], and UltraEdit [Zhao *et al.*, 2024] curate large datasets of source images, target images, and edit instructions, allowing diffusion models to incorporate editing prompts directly during training. Flux Kontext [Batifol *et al.*, 2025] extends this strategy with a rectified flow transformer and concatenates text tokens from the instruction prompt with image tokens from the source image, thus allowing mutual interaction during the joint-attention operation. Finally, some methods adapt the model at edit time to better align with a specific source image. Unitune [Valevski *et al.*, 2023] fine-tunes the noise predictor on the source image and accompanying text description, and then applies SDEdit for editing. Imagic [Kawar *et al.*, 2023] optimizes both the prompt embedding and the model weights, enabling smooth interpolation between the source and target prompts. Training-based methods often achieve stronger semantic control and sometimes higher edit fidelity than training-free methods, but require additional training, heavier computation and rely on efficient data collection or carefully designed strategies to construct pairs of source images, edited images, and editing instructions, which further limits their scalability. Our method overcomes this by fine-tuning on a single image, ensuring a well-integrated edit.

Deep learning in forensics. Deep learning has also been applied to forensic workflows. Notably, [Majumdar *et al.*, 2021; Majumdar *et al.*, 2019] adapt facial recognition networks to

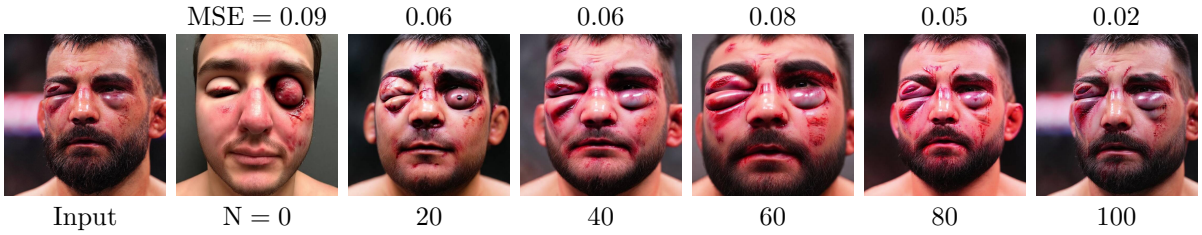


Figure 2: Single-image fine-tuning improves inversion performance.

injured faces, enabling embedding-based matching against reference databases. While these methods operate at the embedding level for algorithmic matching, FlowID instead modifies the image itself, producing a reconstruction suitable for recognition by next-of-kin.

3 Method

3.1 Background

Latent Flow-Matching Models. Modern text-to-image flow matching models generate samples from a data distribution p_0 by iteratively denoising a Gaussian noise sample z_t. This process is described by the *probability flow ODE*:

$$\frac{dz_t}{dt} = v(z_t, t), \tag{1}$$

where the right-hand term, called the *velocity*, can be learned via the conditional flow matching objective:

$$\mathcal{L}_{CFM} = \mathbb{E}_{t,z_0,\epsilon}\|v_\theta(z_t,t) - (\dot{\alpha}_t z_0 + \dot{\sigma}_t \epsilon)\|_2^2 \tag{2}$$

where v_θ denotes the learned velocity, $\epsilon \sim \mathcal{N}(0, I_d)$, and $z_t = \alpha_t z_0 + \sigma_t \epsilon$ is an interpolation between ϵ and z_0. For the linear interpolant $\alpha_t = 1 - t$ and $\sigma_t = t$, which is the most commonly used, the target velocity simplifies to $\epsilon - z_0$. The aforementioned processes are carried out in the latent space of a VAE with encoder $\mathcal{E}$ and decoder $\mathcal{D}$, such that for a given image I, $z_0 = \mathcal{E}(I)$ and $I = \mathcal{D}(z_0)$. Additionally, the velocity prediction $v_\theta(z_t, t)$ can be influenced by supplementary conditions such as a text prompt, which we denote as P.

Image Inversion. A common strategy for image editing with text-conditioned flow matching models relies on inversion. In this setting, one can solve (1) forward in time, starting from $t = 0$, with the source latent z_0^{src} and the source description P^{src}, up to a timestep t_s, referred to as the *strength*. Then, editing can be achieved by solving (1) again, in reverse time, starting from z_{t_s} and using the target textual conditioning P^{tgt} to obtain the edited latent z_0^{tgt}.

3.2 Single-Image Fine-Tuning

Editing large regions using inversion-based methods requires integrating Equation (1) up to a large t_s. As previously described, this poses the problem of error accumulation when using numerical ODE solvers for the inversion. If the inversion is imperfect, crucial identity information is lost and cannot be recovered during reconstruction. We argue that part of the error induced in the inversion stems from the fact that the image to edit is not a sample generated by the model and the semantic alignment between the inversion prompt and the image is suboptimal. Our method begins with single-image fine-tuning on the user-supplied image and prompt, using the pretrained Stable Diffusion 3 [Esser *et al.*, 2024] flow-matching model. This procedure adjusts the model such that the provided image becomes a more likely sample under the conditioning prompt P^{src}. Technically, this amounts to applying the usual flow-matching objective (Eq. 2) to just one image–prompt pair.

Figure 2 shows the effect of single-image fine-tuning on inversion quality, measured by the MSE between the input and reconstructed images. We perform full inversion on the original image, stopping slightly before reaching pure noise (2 steps), in order to retain structural information. As the number of fine-tuning steps N increases, reconstruction error decreases steadily. Fine-tuning progressively recovers global structure, with head pose and identity-related attributes such as facial hair emerging early in the process.

3.3 Localized Editing with Inferred Masks

Recognizing a person from an image is based on visible morphological characteristics, *e.g* the shape of the nose, mouth, jaw, or the distance between the eyes. However, recognition can also depend on secondary cues such as tattoos or jewelry. In practice, these fine-grained details are often not faithfully preserved by the inversion and generation process, and may be inadvertently altered during reconstruction and thus hinder recognition. To address these limitations, our approach incorporates *localized editing*, which explicitly preserves regions of the image that should remain unchanged.

Deciding which areas to preserve and which to modify is a non-trivial challenge, particularly when relying only on textual prompts. To overcome this, we exploit internal signals from the model itself. FlowID leverages the attention maps of the generative model to construct the mask M.

Prior work has shown that the intermediate layers of transformer blocks capture rich semantic information [Tumanyan *et al.*, 2023; Luo *et al.*, 2023; Epstein *et al.*, 2023; Helbling *et al.*, 2025], which can be repurposed to guide image editing [Hertz *et al.*, 2022; Parmar *et al.*, 2023; Epstein *et al.*, 2023; Brack *et al.*, 2024].

For every transformer block in the DiT, consider a sequence of image tokens $z \in \mathbb{R}^{h\times w\times d}$ and of text tokens $c \in \mathbb{R}^{l\times d}$. Inside this text tokens sequence is a tokens subsequence corresponding to the concept of interest that we wish to edit, denoted as $c^* \in \mathbb{R}^{k\times d}$, with $k \leq l$.

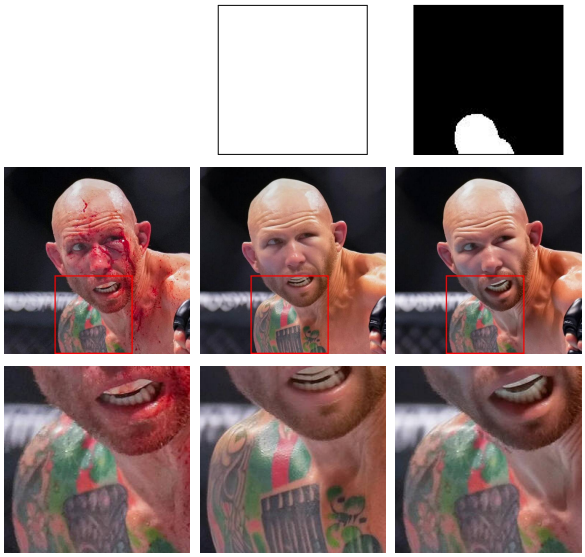


Figure 3: Impact of our masking component for details preservation. Without masking, the person's tattoo is altered *(middle)* When using our masking component, we manage to infer the location of the tattoo and explicitly preserve it *(right)*.

Then, for a given attention head, we extract the query q_z for all the image tokens, as well as the key k_{c^*} for c^*:

$$q_z = Q_z z,\ k_{c^*} = K_c c^* \tag{3}$$

We compute the image concept interaction map $\alpha \in \mathbb{R}^{h \times w \times k}$ as $\alpha = \text{softmax}(q_z k_{c^*}^T)$.

By averaging over concept tokens and reshaping the obtained vector into a matrix, we obtain a cross-attention map $m_{a,b} \in \mathbb{R}^{h \times w}$, for head a and block b, indicating where the concept is present in z. We then average over all attention heads and all transformer blocks, which yields a more salient map $\bar{m}$. Finally, we binarize $\bar{m}$ with a K-means algorithm with 2 barycenters, effectively capturing the binary nature of $\bar{m}$, to produce our final mask $M \in \{0,1\}^{h \times w}$.

Recomputing the mask at every generation step proves ineffective, as the signal tends to vanish toward the end of the backward process. Instead, we compute it once at a specific step t_M, chosen such that $t_M < t_s$, ensuring that the object to be edited is well localized set at the time of mask estimation. For concept removal tasks, this condition is naturally satisfied since the mask can be obtained during inversion. For concept addition, however, we proceed differently: the backward process is first run from t_s down to t_M without masking, the mask is then computed at t_M, and finally the process is backtracked to t_s before resuming the actual editing.

[Couairon *et al.*, 2022] noticed that it is beneficial to have a mask that overshoots the area to be edited. We provide such property to FlowID by applying a max pooling operation to make it bigger.

Figure 3 illustrates the effect of our masking strategy when removing blood while preserving a body tattoo. Without masking, the generative process alters fine tattoo details, reducing recognition reliability. In contrast, explicit masking prevents unintended modifications by enforcing exact reconstruction within the tattoo region.

Algorithm 1 FlowID algorithm

Input: $v_\theta, z_0^{src}, c^{src}, c^{tgt}, c^*, t_s, \{\sigma_t\}_t$
Output: z_0^{tgt}

$v_\theta \leftarrow \text{FineTune}(v_\theta, z_0^{src}, c^{src})$
$z_t \leftarrow z_0^{src}$
$z_0^{ref} \leftarrow z_0^{src}$
for $t = 0, \ldots, t_{s-1}$ **do**
  $z_{t+1} \leftarrow z_t + (\sigma_{t+1} - \sigma_t) v_\theta(z_t, t, c^{src})$
  $z_{t+1}^{ref} \leftarrow z_{t+1}$
end for
$M \leftarrow \text{GetMask}(\{z_t^{ref}\}_t, c^{src}, c^{tgt}, c^*)$
for $t = t_s, \ldots, t_1$ **do**
  $z'_{t-1} \leftarrow z_t + (\sigma_{t-1} - \sigma_t) v_\theta(z_t, t, c^{tgt})$
  $z_{t-1} \leftarrow M \cdot z'_{t-1} + (1 - M) \cdot z_{t-1}^{ref}$
end for
$z_0^{tgt} \leftarrow z_t$
return z_0^{tgt}

To summarize, let us consider the following reference inversion trajectory under v_θ and a given encoded prompt c^{src} : $\{z_0^{ref}, z_{t_1}^{ref}, \ldots, z_{t_s}^{ref}\}$. Assuming we have a binary mask M that separates areas that should be edited from areas that must remain untouched, we paste reference parts on the current ODE solver iterate z'_{t-1}: $z_{t-1} = M * z'_{t-1} + (1 - M) * z_{t-1}^{ref}$ in a similar fashion to [Couairon *et al.*, 2022]. A pseudocode for the FlowID method can be seen in Algorithm 1.

4 Experiments

In this section, we present a comprehensive experimental evaluation of our approach. We first introduce InjuredFaces, a new benchmark for facial reconstruction under severe injuries, explicitly designed to evaluate identity preservation. We then evaluate FlowID using Stable Diffusion 3 and compare it against recent state-of-the-art editing methods, including UltraEdit, SDEdit, ICEdit, Kontext, and RF-Solver. To assess generalization beyond injury reconstruction, we further evaluate our method on classical and sequential facial editing tasks using a subset of the FFHQ dataset [Karras *et al.*, 2019]. Finally, we present an ablation study to quantify the contribution of each component of our approach. Additional hyperparameters and implementation details are provided in Appendix A.

4.1 The InjuredFaces Benchmark

Evaluating facial reconstruction quality is essential for forensic identification, yet reproducible evaluation is hindered by the impossibility of sharing real forensic images due to privacy and ethical constraints. To overcome this barrier, we introduce InjuredFaces, a new benchmark for generative facial reconstruction under severe injuries, built from publicly available images of professional athletes. InjuredFaces is a core contribution of this work and is intended as a standardized, reusable resource for the community to study identity-preserving reconstruction in extreme conditions.

InjuredFaces consists of 755 injured facial portraits, denoted I_{inj}, corresponding to $N = 449$ distinct identities. For each identity, we additionally provide a set of uninjured reference portraits, denoted I_{ref}, resulting in a total of 5,213 reference images across the dataset. These uninjured portraits serve as identity anchors and allow the construction of a stable reference representation for each individual.

For each injured image, we associate a formatted textual prompt P. The prompt combines a fixed identity-neutral description, "a person's face", with a concise textual description of the visible injuries (e.g., "open wound on the forehead"). This formulation enables controlled editing while keeping the identity description consistent across samples.

To represent identity information, we rely on a pretrained face recognition embedding network. Following prior work, we use `fal/AuraFace-v1` as our identity encoder, denoted $\mathcal{A}$. Given an image I, the network maps it to a fixed-dimensional latent embedding $\mathcal{A}(I)$ that captures identity-related facial characteristics. For each identity, we compute a reference embedding a by averaging the embeddings of its corresponding uninjured portraits in I_{ref}. These embeddings provide a consistent identity representation that we later use to assess identity preservation after reconstruction. Representative examples from InjuredFaces are shown in Appendix B.

Facial reconstruction quality is assessed along five complementary dimensions: identity preservation, injury removal effectiveness, edit faithfulness, image quality, and perceptual similarity.

Identity Preservation is assessed using the identity embeddings defined above. Given a reconstructed image and its corresponding reference identity embedding, we measure preservation by computing the cosine similarity between their embeddings. This score reflects how well identity-related facial characteristics are retained after reconstruction. Examples of identity preservation values computed over representative image pairs, as well as the resulting score distributions, are reported in Appendix B.

Injury Removal is assessed using a score derived from a large Vision–Language Model (VLM). Prior work has shown that large VLMs can be reliably repurposed for binary image classification tasks [Kumari *et al.*, 2025]. Given an edited image, we query the VLM with the question *"Is this person's face injured? Only answer with yes or no."* We then extract the logits corresponding to the *Yes* and *No* tokens and convert them into a scalar score using a sigmoid function: $\text{VLMScore} = \sigma(l_{\text{Yes}} - l_{\text{No}})$. Lower scores indicate more successful injury removal. This metric provides a robust, semantically grounded measure of injury presence.

Edit Faithfulness is measured using the CLIP score, computed as the cosine similarity between the CLIP embedding of the edited image and that of the target prompt. This measure captures how well the edited image aligns with the intended textual description.

Image Quality is evaluated in terms of visual realism and distributional consistency using FID [Heusel *et al.*, 2017] and CMMD [Yang *et al.*, 2024] (which has been shown to better correlate with human judgments of image quality). Both metrics are computed with respect to a reference set of 14,000 images from FFHQ.

Perceptual Similarity is quantified using LPIPS [Zhang *et al.*, 2018] between the original and edited images. This metric complements identity preservation by capturing low-level visual deviations introduced by the editing process.

To ensure a fair evaluation, IP, FID, CMMD, and LPIPS are computed only on samples where the edit is successful, defined as $\text{VLMScore} < 0.5$. This prevents trivial preservation scores resulting from failed edits. The proportion of successful edits is given in Appendix A.

4.2 Results

Qualitatively, Figure 4 presents a comparison for different individuals from the InjuredFaces benchmark. FlowID consistently achieves a better balance between injury removal and identity preservation. It removes severe facial injuries while introducing minimal unintended changes to the original appearance. Methods relying on heavier backbones, such as Kontext and RF-Solver, often succeed at removing visible injuries but struggle to preserve identity. For instance, in the first and fourth rows, RF-Solver removes the targeted injury but alters global facial structure, leading to noticeable identity drift. In addition, the tattoo close-up in the last row shows that several baselines inadvertently deform identity-relevant details, whereas FlowID better preserves such fine-grained cues. SDEdit applies strong edits but suffers from substantial identity degradation due to its non-inversion-based formulation. UltraEdit and ICEdit, although effective on in-distribution editing tasks, perform poorly in this setting. Both methods are trained on curated editing pairs and struggle with the out-of-distribution nature of severe facial trauma. ICEdit, despite relying on a strong Flux-based backbone, remains constrained by its training regime and lacks robustness for unseen forensic edits.

Quantitatively, Table 1(a) reports results on the full InjuredFaces benchmark. FlowID achieves the highest identity preservation score among all evaluated methods, demonstrating superior retention of identity-critical facial features after reconstruction. This result is particularly significant given the severity of the edits required for injury removal. At the same time, FlowID delivers markedly better injury removal performance. It attains the lowest (best) VLM score and the best CLIP alignment by a wide margin, indicating both effective suppression of injury-related artifacts and strong faithfulness to the target prompt. FlowID ranks second in terms of LPIPS, behind Kontext; this difference is consistent with its substantially higher injury removal success, as stronger and more complete edits necessarily induce larger perceptual changes. Importantly, these improvements do not sacrifice identity fidelity, in contrast to competing approaches. FlowID further maintains FID and CMMD on par with those of other methods. Beyond metric improvements, these results are directly relevant to forensic identification. Reliable suppression of injury-related artifacts, combined with strong identity preservation, increases the likelihood that reconstructed images remain recognizable by next-of-kin while reducing unnecessary visual distress.

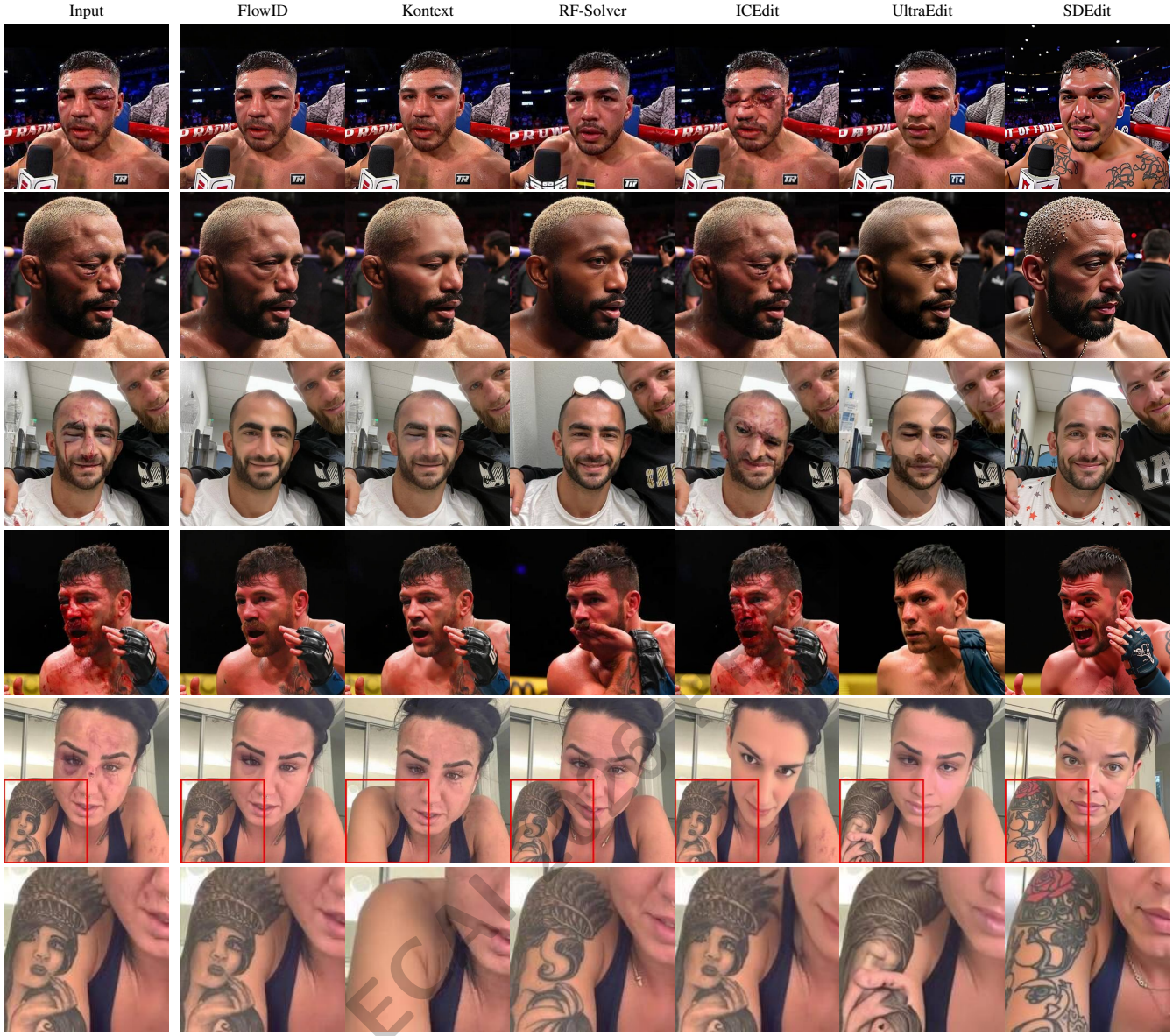


Figure 4: Qualitative comparison of image editing methods on InjuredFaces.

4.3 Additional Results on FFHQ

To evaluate the capacity of FlowID to generalise beyond injury reconstruction, we report additional results on a subset of the FFHQ dataset focusing on classical facial attribute editing such as hair color, beard, glasses, and facial expression. For each image, we apply a single textual edit instruction while keeping the original image as reference. This setting allows us to assess whether FlowID maintains strong identity preservation and localized control in standard, non-forensic editing scenarios. Table 1(b) reports results on single-shot edits. In this setting, FlowID achieves the highest identity preservation score among all methods, while maintaining VLM edit score (higher is better) comparable to Kontext, indicating strong generalization beyond injury-specific edits. In addition, FlowID attains the lowest FID, CMMD, and LPIPS values, indicating superior visual fidelity and minimal unintended perceptual changes under successful edits.

Qualitative examples and additional results on sequential editing can be found in our project page[2].

4.4 Ablation Study

To assess the contribution of each component, we perform an ablation study on FFHQ by removing fine-tuning and masking individually. Results are reported in Table 2.

Without fine-tuning, identity preservation drops substantially (0.52 vs. 0.77), and CMMD increases (1.08 vs. 0.45), indicating weaker semantic and distributional alignment. Nevertheless, even without fine-tuning, the method already matches

[2]https://jrpll.github.io/flowid/

Method	IP ↑	CLIP ↓	FID ↓	CMMD ↓	LPIPS ↓	VLM ↓
UltraEdit	0.30	0.23	116.44	2.36	0.20	0.19
SDEdit	0.23	0.24	**104.48**	**2.29**	0.24	0.15
ICEdit	0.22	0.25	190.43	2.61	0.24	0.88
Kontext	0.48	0.24	128.78	3.47	**0.10**	0.28
RF-Solver	0.48	0.24	117.88	3.12	0.14	0.36
FlowID	**0.50**	**0.23**	123.07	3.09	0.12	**0.14**

(a) InjuredFaces

Method	IP ↑	CLIP ↑	FID ↓	CMMD ↓	LPIPS ↓	VLM ↑
UltraEdit	0.46	0.26	35.28	0.83	0.24	0.84
SDEdit	0.13	0.28	69.10	1.80	0.49	0.61
ICEdit	0.75	0.27	36.37	0.58	0.19	0.65
Kontext	0.72	0.27	31.30	0.85	0.20	**0.96**
RF-Solver	0.26	**0.28**	51.59	1.38	0.31	0.87
FlowID	**0.77**	0.27	**29.04**	**0.45**	**0.11**	0.94

(b) FFHQ

Table 1: Quantitative comparison on (a) InjuredFaces and (b) FFHQ.

Method	IP ↑	CLIP ↑	FID ↓	CMMD ↓	LPIPS ↓	VLM ↑
w/o fine-tuning	0.52	0.27	32.32	1.08	0.15	**0.99**
w/o masking	0.61	0.27	34.95	0.63	0.24	0.96
FlowID	**0.77**	0.27	**29.04**	**0.45**	**0.11**	0.94

Table 2: Ablation on fine-tuning and masking components.

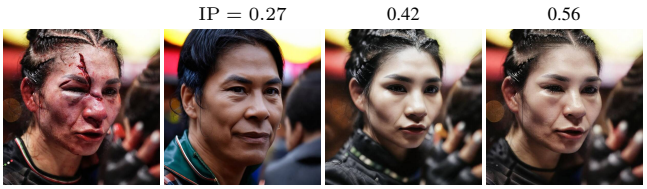


Figure 5: Edit performance as a function of fine-tuning steps.

or exceeds several competing approaches in identity preservation on FFHQ (e.g., outperforming UltraEdit at 0.46 and SDEdit at 0.13).

The effect of single-image fine-tuning is further illustrated in Figure 5. As the number of fine-tuning steps increases, identity preservation improves steadily, rising from 0.27 without fine-tuning to 0.56 after 100 steps. Qualitatively, fine-tuning progressively restores identity-defining facial traits such as facial structure and proportions, while preserving the effectiveness of injury removal.

Removing masking degrades realism and perceptual quality, with FID increasing from 29.04 to 34.95 and LPIPS from 0.11 to 0.24, confirming that unconstrained edits introduce unnecessary visual distortions. Despite this degradation, the unmasked variant still achieves stronger identity preservation (0.61) than most baselines, including UltraEdit and SDEdit, while maintaining a high editing success rate (VLM 0.96).

Figure 3 illustrates the effect of localized masking. Without masking, the model alters regions unrelated to the injury, degrading fine-grained identity cues such as tattoos and increasing perceptual distortion, whereas applying the inferred mask enforces exact reconstruction outside the editable region, leading to improved LPIPS and FID while preserving identity-critical details.

When combined, fine-tuning and masking yield consistent improvements across all metrics, achieving the best trade-off between identity preservation, image quality, and edit success. The slight decrease in success rate with masking reflects a deliberate trade-off: restricting edits to semantically relevant regions limits prompt flexibility but results in higher fidelity and perceptual quality.

5 Conclusion

In many medico-legal and humanitarian identification workflows, facial imagery is often the only usable visual evidence. Severe trauma, post-mortem degradation, and occlusions can remove key facial cues, slowing identification, increasing forensic workload, and prolonging uncertainty for families. This paper re-frames facial reconstruction as a recognition focused task that prioritizes identity preservation and strict edit containment, and makes two primary contributions.

First, we introduce FlowID, an identity preserving facial editing framework that restores recognizable cues while constraining unintended changes. FlowID combines per-image fine-tuning, adapting the generator to each image, with an attention-derived masking mechanism that localizes edits to damaged regions while enforcing exact reconstruction elsewhere. Second, we propose InjuredFaces, a benchmark and evaluation protocol for reconstruction under severe facial trauma, where identity preservation is the primary objective.

FlowID is validated experimentally and has already shown operational impact. It has contributed to successful identifications, including those of deceased migrants recovered in the Mediterranean, and has been deployed across several medico-legal sites, in Mexico and Colombia. These deployments highlight the practical relevance of identity centric reconstruction in reducing processing time and cognitive burden in resource constrained institutions, and may also help limit repeated exposure to graphic imagery.

To support reproducible progress, InjuredFaces provides a standardized benchmark aligned with realistic forensic conditions. Unlike evaluations focused on generic realism or stylistic fidelity, it centers around identity preservation, injury removal, and unintended modifications, encouraging methods better matched to high stakes identification workflows.

FlowID has limitations. Per-image fine-tuning adds computational overhead, though it remains lightweight, can be reduced through parameter efficient adaptation such as LoRA style updates, and amortizes over sequential edits since later modifications run at standard generation cost. Furthermore, our masking component can prove less useful when injuries are globally present on the face. Improved masking integration and lower cost adaptation are our future work.

Method	FFHQ					InjuredFaces
	Step 1	Step 2	Step 3	Step 4	Step 5	Step 1
UltraEdit	0.85	0.84	0.83	0.80	0.78	0.84
SDEdit	0.58	0.58	0.57	0.56	0.56	0.86
ICEdit	0.64	0.58	0.54	0.46	0.43	0.10
Kontext	0.97	0.96	**0.96**	0.93	0.92	0.52
RF-Solver	0.88	0.86	0.86	0.85	0.85	0.67
FlowID	**0.97**	**0.96**	0.95	**0.94**	**0.93**	**0.95**

Table 3: Successful edit rates per method and benchmark.

Figure 6: An InjuredFaces sample.

A Implementation Details

Both InjuredFaces and FFHQ benchmarks were run on a A100 GPU cluster. For methods that use source and target prompts like ours, we use the prompt describing the injured photo from InjuredFaces as the source, and *"a person's face"* as the target. For instruction-based methods, we infer the edit instructions based on the source and target prompt with simple rules. For implementations that allow negative guidance, we extract the injury description from the source prompt by removing *"a person's face"* from the prompt.

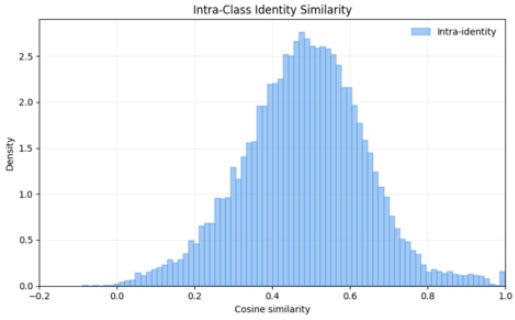
Our method uses Adam8Bit optimizer in order to lower the memory requirements for compatibility with consumer hardware. We run all fine-tunings with batch size of 1, gradient accumulation of 10, 80 optimizer steps and use a learning rate of 5e-5. These details allow us to run FlowID on a NVIDIA RTX4090 GPU.

For instruction-based methods we check for the presence of a given type of injury and turn it into an instruction, *e.g*, if "wound" is in the description, the instruction becomes "remove the wound". Checked attributes are : "wound", "blood", "mouth open", "bump", "bruise", "cut"", "injury", "swollen". For ICEdit, we used the following prompt : "A diptych with two side-by-side images of the same scene. On the right, the scene is exactly the same as on the left but instruction" where the instruction with the above rule. For RF-Solver, we use default parameters except for the number of features injection steps, which we set to 10, as it yielded the best balance between identity preservation and edit insertion. Similarly we used a strength t_s of 0.65 for SDEdit.

Success rates for both sequential edits on FFHQ and one-step edits on InjuredFaces can be seen in Table 3. We notice that our method tends to edit more strongly than others, thus motivating the need for the filtering mentioned in Section 4.1.

B Additional Details on InjuredFaces

Figure 6 shows a sample from our InjuredFaces dataset. Each injured faces represents a row in our dataset. It comes with healthy portraits of the same person, a textual description of the injured portraits, and an embedding vector a, computed as the average of identity embeddings.



(a) Intra-class identity similarity distribution.

(b) Cosine similarity matrix between identity embeddings.

Figure 7: Identity similarity analysis.

The histogram shown in Figure 7a reports the intra-class identity similarity distribution computed over reference images, showing a stable and well-shaped distribution centered around moderate to high similarity values.

Figure 7b provides additional insight into the identity preservation metric. The cosine similarity matrix illustrates pairwise identity similarities across multiple identities, with higher values along same-identity pairs and substantially lower values for mismatched identities.

Together, these visualizations confirm that the identity embedding produces consistent representations for the same individual, supporting its use as a reference for identity preservation evaluation.

Ethical Statement

This work raises important questions regarding the processing of images of deceased individuals without their consent. We acknowledge this tension, but note that it presents an inherent dilemma: the very purpose of altering these images is to identify the deceased and return them to their families. In the absence of other clues, involving communities through care-

fully processed images may be the only viable path toward recognition. To ensure responsible deployment, all processing is conducted locally by authorities responsible for identification, and reconstructed images are shared only through official channels, clearly marked as renderings. We also recognize that reconstructions may mislead family members; our system is therefore designed to support—not replace—existing identification workflows, with final determinations always involving verification by forensic professionals.

Acknowledgements

This work was supported by a French government grant managed by the Agence Nationale de la Recherche under the "Investissements d'avenir" program (reference "ANR-21-ESRE-0051").

References

[Batifol *et al.*, 2025] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025.

[Brack *et al.*, 2024] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8861–8870, 2024.

[Brooks *et al.*, 2023] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023.

[Couairon *et al.*, 2022] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance, 2022.

[Epstein *et al.*, 2023] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023.

[Esser *et al.*, 2024] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[Helbling *et al.*, 2025] Alec Helbling, Tuna Han Salih Meral, Ben Hoover, Pinar Yanardag, and Duen Horng Chau. Conceptattention: Diffusion transformers learn highly interpretable features. *ICML 2025: Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[Hertz *et al.*, 2022] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control.(2022). *URL https://arxiv. org/abs/2208.01626*, 1, 2022.

[Heusel *et al.*, 2017] 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.

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

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[Huberman-Spiegelglas *et al.*, 2024] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12469–12478, 2024.

[Kadosh *et al.*, 2025] Edo Kadosh, Nir Goren, Or Patashnik, Daniel Garibi, and Daniel Cohen-Or. Tight inversion: Image conditioned inversion for real image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6980–6990, 2025.

[Karras *et al.*, 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[Kawar *et al.*, 2023] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6007–6017, 2023.

[Kumari *et al.*, 2025] Nupur Kumari, Sheng-Yu Wang, Nanxuan Zhao, Yotam Nitzan, Yuheng Li, Krishna Kumar Singh, Richard Zhang, Eli Shechtman, Jun-Yan Zhu, and Xun Huang. Learning an image editing model without image editing pairs. *arXiv preprint arXiv:2510.14978*, 2025.

[Labs, 2024] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

[Lipman *et al.*, 2023] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[Liu *et al.*, 2022] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[Luo *et al.*, 2023] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic

correspondence. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[Majumdar *et al.*, 2019] Puspita Majumdar, Saheb Chhabra, Richa Singh, and Mayank Vatsa. Subclass contrastive loss for injured face recognition. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7. IEEE, 2019.

[Majumdar *et al.*, 2021] Puspita Majumdar, Saheb Chhabra, Richa Singh, and Mayank Vatsa. Recognizing injured faces via scifi loss. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1):112–123, 2021.

[Meng *et al.*, 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.

[Mokady *et al.*, 2023] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023.

[Parmar *et al.*, 2023] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023.

[Peebles and Xie, 2023] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, October 2023.

[Sheynin *et al.*, 2024] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024.

[Sohl-Dickstein *et al.*, 2015] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

[Song *et al.*, 2021a] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[Song *et al.*, 2021b] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[Tumanyan *et al.*, 2023] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, June 2023.

[Valevski *et al.*, 2023] Dani Valevski, Matan Kalman, Eyal Molad, Eyal Segalis, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning a diffusion model on a single image. *ACM Trans. Graph.*, 42(4), July 2023.

[Wang *et al.*, 2024] Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024.

[Yang *et al.*, 2024] Ruihan Yang, Hannes Gamper, and Sebastian Braun. Cmmd: Contrastive multi-modal diffusion for video-audio conditional modeling. In *European Conference on Computer Vision*, pages 214–226. Springer, 2024.

[Zhang *et al.*, 2018] 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*, pages 586–595, 2018.

[Zhang *et al.*, 2025] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*, 2025.

[Zhao *et al.*, 2024] Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024.