SAFE-FM: Secure, Anonymous Face Editing using Flow Matching

Aarav Gang[1], Samarth Bhatia[1], Anagha Pillalamarri[1],Siddharth Bayapureddy[1]

[1]Birla Institute of Technology and Science, Pilani, Hyderabad Campus

Motivation & Novelty

The Privacy-Utility Dilemma In the realm of digital privacy, face anonymization has traditionally forced a compromising trade-off between concealing identity and preserving utility. Current methods fail to balance these needs effectively:

- **Naive Techniques (Blurring/Pixelation):** While effective at masking identity, these methods destroy non-verbal communication cues, such as smiles, gazes, and micro-expressions, rendering the video unnatural and unengaging.
- **Generative Diffusion Models:** Advanced models like Stable Diffusion can synthesize photorealistic faces but suffer from high computational latency. Their reliance on iterative denoising (20–50 steps per frame) makes them fundamentally unsuitable for real-time applications like live video conferencing.

Our Solution: Flow Matching We propose **SAFE-FM**, a novel framework that bridges this gap by replacing the slow, stochastic denoising of diffusion models with a fast, deterministic Flow-Matching backbone. Unlike diffusion, Flow Matching learns a continuous velocity field that maps noise to data via a deterministic Ordinary Differential Equation (ODE). This allows us to reduce the inference process to a mere 3–5 steps without compromising visual fidelity, finally unlocking the capability for real-time, expression-preserving anonymization.

Primary Objectives

The primary objective of this research is to engineer the first real-time face anonymizer capable of operating on standard consumer hardware. Our key goals include:

- **Real-Time Performance:** Achieve inference speeds of $\geq$25 FPS on live webcam feeds, ensuring seamless integration with video streaming platforms.
- **Expression Fidelity:** Decouple identity from motion to ensure that while the subject's facial features are replaced, their pose, lip movements, and emotional expressions remain synchronized with the original feed.
- **Privacy Guarantee:** Generate a consistent, artificial identity that retains no reversible mathematical link to the source face, effectively preventing reconstruction attacks.
- **Technical Superiority:** Demonstrate the advantage of deterministic velocity integration (Flow Matching) over stochastic diffusion models in terms of speed and stability.

System Architecture

The Anonymization Pipeline Our system processes live video through a multi-stage generative pipeline designed for speed and consistency:

Landmark Extraction: We detect 3D facial keypoints (capturing the eyes, eyebrows, lips, and jawline) to generate dense expression heatmaps.

Conditioning: These heatmaps act as control signals and are injected into our Flow-Matching Generator, a fine-tuned UNet architecture.

Identity Injection: To ensure the output is not just a reconstruction of the original user, we inject a constant Identity Embedding, which forces the model to synthesize a specific, pre-defined artificial identity.

Temporal Warm-Start: To prevent "flickering" artifacts, we employ a warm-start strategy, initializing the generation of the current frame using the latent representation of the previous frame.

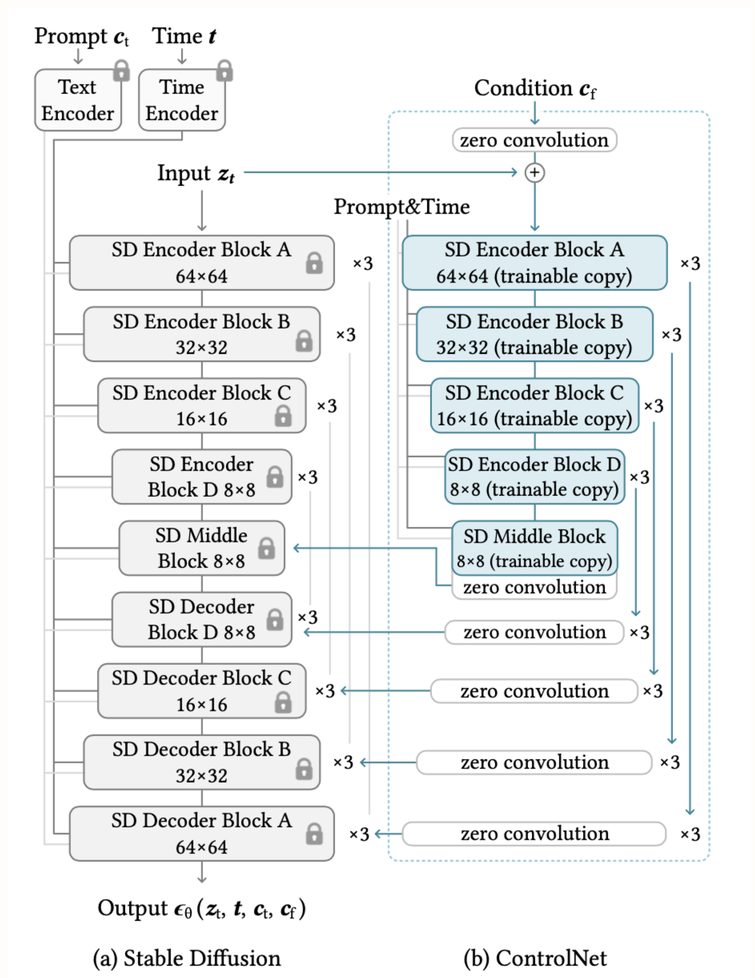

Figure 1: **Modified Architecture:** We adapt the ControlNet structure to a Flow-Matching backbone. The locked gray blocks represent the pre-trained model, while the trainable blue blocks inject landmark conditions via "Zero Convolution" layers.

Fine Tuning Strategy

To adapt a massive pre-trained generative model without prohibitive computational costs, we utilize Parameter-Efficient Fine-Tuning (PEFT). This approach allows us to specialize the model for face generation while keeping it lightweight:

- **Low-Rank Adaptation (LoRA):** We freeze the pre-trained backbone and inject small, trainable rank-decomposition matrices into the attention layers. This allows us to fine-tune the model using only ~1% of the total parameters.
- **ControlNet Adapter:** Simultaneously, we train a lightweight side-network that processes the landmark heatmaps into multi-scale features. This adapter guides the main UNet to strictly adhere to the spatial constraints of the user's expression, ensuring that if the user smiles or blinks, the anonymized face does exactly the same.
- **Temporal Warm-Start:** To ensure video consistency and prevent inter-frame flickering, we implement a temporal warm-start strategy during inference. Instead of initializing the flow generation from random Gaussian noise for every frame, we initialize the state using the latent representation of the previous anonymized frame. This enforces temporal coherence without requiring computationally expensive video-specific training.

Figure 2: Visualizing Spatial Control: Just as ControlNet (shown above) uses pose skeletons to dictate the structure of a generated image, SAFE-FM uses facial landmarks to force the anonymized output to strictly mimic the source subject's pose and expression.

Core Hypothesis: Why Flow Matching?

Determinism vs. Stochasticity The core hypothesis of this work is that Flow Matching serves as a superior alternative to Diffusion for real-time tasks due to its mathematical formulation.

- **Diffusion Models:** Rely on a stochastic reverse SDE, requiring a long trajectory (20–50 steps) to resolve an image from noise.
- **Flow Matching (Ours):** Learns a continuous velocity field $v_t(x)$ that maps the noise distribution p_0 directly to the data distribution p_1 via a deterministic Ordinary Differential Equation (ODE):

$$dX_t = v_t(X_t)dt \tag{1}$$

Objective Function We train the model by regressing this velocity field directly, which allows for straighter generation trajectories and faster convergence:

$$\mathcal{L}_{FM} = \mathbb{E}_{t,x_1,x_0}||v_t(x_t) - (x_1 - x_0)||^2 \tag{2}$$

This linearity allows us to use efficient numerical integrators (like Euler or RK45) to traverse the generative path in just **3–5 steps**, enabling real-time performance.

Performance Comparison

We project a significant performance leap over traditional diffusion-based anonymizers. By shifting to a deterministic Flow-Matching backbone, we achieve real-time speeds without sacrificing generative quality.

Metric	Diffusion	SAFE-FM (Ours)
Math	Stochastic	**Deterministic**
Steps	20–50	**3–5**
Speed	Slow (Secs/Frame)	**Real-Time (25+ FPS)**
Stability	Low (Flickers)	**High (Stable)**

References

[1] Lipman, Y. et al. (2023). Flow Matching for Generative Modeling. *ICLR*.

[2] Rombach, R. et al. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. *CVPR*.

[3] Zhang, L. et al. (2023). Adding Conditional Control to Text-to-Image Diffusion Models. *ICCV*.

[4] Hu, E.J. et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*.