FLOWFACESWAP: TRAINING-FREE VIDEO FACE SWAPPING VIA STRUCTURE-PRESERVED IDENTITY FLOW EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

High-fidelity video face swapping requires transferring a source identity seamlessly while strictly preserving the target's expressions, poses, and temporal consistency. Achieving satisfactory swapping performance typically necessitates resource-intensive, task-specific training, making these models expensive and inherently limiting their generalization to unseen scenarios. Recent training-free approaches have emerged as promising alternatives and achieved notable progress. However, they still struggle with temporal inconsistencies and identity leakage in face swapping. In this work, we reformulate the face swapping task as an identity-guided flow residual generation process, and propose a training-free framework FlowFaceSwap based upon a frozen rectified flow prior. Specifically, by optimizing compact identity tokens at test time, our method establishes a discriminative identity context. To prevent the overwriting of target dynamics, we introduce a structure-aware residual modulation mechanism driven by temporally tracked semantics and landmark-guided gates. This spatial control accurately localizes identity transfer while rigorously maintaining the target's subtle expressions and fine-grained motions. Extensive experiments on FaceForensics++ and CelebVHQ demonstrate that FlowFaceSwap generates temporally stable, highly realistic results. It significantly outperforms existing training-free baselines in both visual quality and identity retention, achieving an overall performance highly competitive with specialized, training-based models.

1 INTRODUCTION

Video face swapping aims to seamlessly transfer a source identity into a target video while rigorously preserving the target structure including expressions, poses, and temporal coherence. This task is inherently challenging due to the severe entanglement between identity and facial structure.

Existing solutions primarily fall into training-based paradigms, with 3D, GAN, or diffusion frameworks (Bitouk et al., 2008; Wang et al., 2021; Lee et al., 2024; He et al., 2025). Specifically, recent diffusion systems achieve impressive realism through dedicated supervision and identity-video joint training (Baliah et al., 2025; Wang et al., 2025; Yu et al., 2024; Kim et al., 2025b; Kang et al., 2025; Guo et al., 2026; Luo et al., 2025; Bondurant et al., 2025). While these systems achieve strong results, their reliance on resource-intensive, task-specific training imposes high computational barriers. As an alternative, training-free personalization methods like SwapAnything (Gu et al., 2025) and PhotoSwap (Gu et al., 2023) offer high flexibility without heavy fine-tuning. Despite their significance, these approaches heavily rely on diffusion inversion techniques (Song et al., 2021; Gal et al., 2023), which are temporally fragile when extended to video generation, suffering from degraded visual quality, severe temporal inconsistency, and noticeable expression and pose misalignment.

To address such inversion limitations, recent inversion-free methods (Liu et al., 2023a; Esser et al., 2024; Lipman et al., 2023; Wan et al., 2025) have emerged as a promising paradigm for efficient flow editing (Kulikov et al., 2025; Li et al., 2025). Building on this generic trajectory-editing paradigm, we study how to adapt velocity-difference editing to reference-specific video face swapping. Directly applying this paradigm poses two critical challenges. 1) *Identity Injection*. Achieving high-fidelity identity transfer in a frozen flow model without relying on resource-heavy fine-tuning

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.

remains an open problem. And 2) *Spatial Awareness during Flow Editing*. Naive flow modification tends to apply latent shifts globally. When applied to face swapping, this uniform shift blindly overwrites delicate target structures, distorting fine-grained facial dynamics and degrading overall temporal stability.

To tackle these challenges, we propose **FlowFaceSwap**, a training-free video face-swapping framework built upon a frozen rectified flow prior. Our approach ensures high-fidelity synthesis without tuning any model weights through three core designs: **1) Test-Time Identity Token Optimization.** For robust identity injection, we introduce a test-time optimization strategy for compact identity tokens. Specifically, we optimize these tokens within the frozen rectified flow model using an identity loss, effectively establishing a highly discriminative identity context for both the source and target subjects. These tokens are optimized once per identity and can be reused across multiple target videos. **2) Identity-Guided Flow Residual Editing.** Rather than performing diffusion inversion, we use optimized source and target identity tokens to steer the flow trajectory. At each sampling step, the difference between the source-conditioned and target-conditioned velocity predictions provides an identity residual, which is subsequently regulated by the semantic region control and landmark-guided target-structure preservation mechanisms. **3) Structure-Aware Residual Modulation.** To resolve the spatial awareness challenge and prevent structural distortions, we introduce a dual-level spatial control mechanism to regulate the applied residual. First, semantic masks are utilized to explicitly control the identity residual, scaling its intensity across different facial regions. Concurrently, a landmark-guided target-structure preservation mechanism is introduced to dynamically suppress the residual in expression regions near the clean endpoint of the sampling trajectory. As illustrated in fig. 1, FlowFaceSwap outperforms existing training-free baselines in both visual quality and identity preservation, achieving an overall performance highly competitive with training-based models. Our main contributions are threefold:

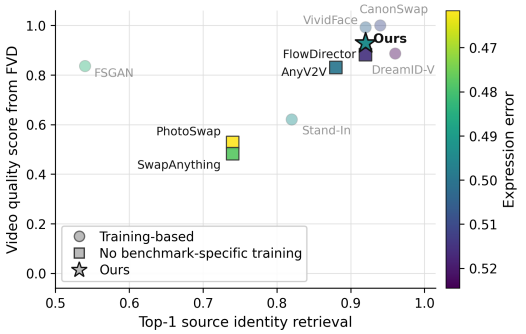


Figure 1: Performance comparison on FF++ (Rössler et al., 2019). FlowFaceSwap outperforms other training-free methods (squares) while achieving highly competitive performance with training-based models (circles).

- We introduce a test-time identity token optimization strategy within frozen rectified flow video priors, thereby establishing robust and discriminative identity context without any task-specific training.
- We adapt inversion-free flow editing to reference-specific video face swapping by deriving an identity-guided residual from optimized source and target identity contexts and applying it directly along the target trajectory, without requiring diffusion inversion.
- We develop a structure-aware residual modulation mechanism driven by temporally tracked semantic masks and landmark-guided target-structure preservation. This localizes identity transfer, ensuring the strict preservation of the target's fine-grained motions and temporal stability.

2 Related Work

Face Swapping. Face swapping has evolved through three main technical paradigms: 3D-centric, GAN-based, and diffusion-driven approaches. Early methods (Blanz & Vetter, 1999; Bitouk et al., 2008; Blanz et al., 2004; Wang et al., 2021) fit explicit 3D Morphable Models (3DMMs) to transfer pose and expression, offering interpretable geometric control but yielding texture-limited results. GAN-based methods (Chen et al., 2020; Lee et al., 2024; Liu et al., 2023b; Zhu et al., 2021) substantially improved visual realism and identity fidelity. Representative models such as SimSwap (Chen et al., 2020) learn identity-preserving generators for image face swapping, while E4S (Liu et al.,

2023b) uses regional GAN inversion to better separate shape and texture. Recent diffusion-based face swapping methods (Baliah et al., 2025; Zhao et al., 2023; Wang et al., 2025; Shao et al., 2025; Yu et al., 2024; Bondurant et al., 2025) improve photorealism and conditional flexibility. Image-level approaches (Kang et al., 2025; Yoo et al., 2025; Han et al., 2024; Kim et al., 2025b), such as DiffFace (Kim et al., 2025b) and FaceAdapter (Han et al., 2024), inject identity through diffusion conditioning, adapters, or recognition features. Video-oriented models further address temporal coherence (Baliah et al., 2026; Wang et al., 2025; Shao et al., 2025; Guo et al., 2026; Luo et al., 2025). VividFace (Shao et al., 2025) uses image-video hybrid training for high-fidelity video swapping, and recent method DreamID-V (Guo et al., 2026) unifies image and video face swapping with explicit identity supervision and strong versatility. However, most of these works rely on resource-intensive, task-specific training processes, limiting their flexibility and scalability in practice.

Subject-Driven Editing and Personalization. Subject-driven generation embeds reference identities into pre-trained models. Textual Inversion (Gal et al., 2023) optimizes a learnable token with a frozen generator, while DreamBooth (Ruiz et al., 2023) fine-tunes model weights for stronger fidelity. Subsequent methods improve efficiency via encoders or adapters (Shi et al., 2024; Ye et al., 2023; Wang et al., 2024; Guo et al., 2024), employing image-prompt adapters (Ye et al., 2023), recognition embeddings (Wang et al., 2024), or contrastive alignment (Guo et al., 2024). Methods for reference-guided replacement and localized personalization are closely related to face swapping by preserving target spatial layouts. Context-aware inpainting (Li et al., 2023; Yang et al., 2023) localizes edits with masks or guidance, while PhotoSwap (Gu et al., 2023) and SwapAnything (Gu et al., 2025) manipulate cross-attention maps to replace subjects while retaining composition. However, these methods are primarily image-based and often rely on diffusion inversion (Song et al., 2021), which is costly and temporally fragile for videos. Inspired by token-learning personalization, our method optimizes source and target tokens at test time for a frozen rectified flow prior, and performs identity residual editing strictly constrained by semantic masks and target landmarks.

Rectified Flow Editing. Editing real images or videos with diffusion models often demands inversion into a latent noise trajectory (Gu et al., 2023; 2025; Ku et al., 2024; Chen et al., 2026), introducing reconstruction errors and high computational costs. Rectified flow and flow-matching models (Liu et al., 2023a; Esser et al., 2024; Lipman et al., 2023) enable direct trajectory editing via their learned velocity fields. For instance, FlowEdit (Kulikov et al., 2025) directly modifies ODE trajectories, FlowAlign (Kim et al., 2025a) and FIA-Edit (Yang et al., 2025) add regularization constraints, and FlowDirector (Li et al., 2025) and FlowMotion (Wang et al., 2026) extend inversion-free flow editing to videos. These methods establish the generic velocity-difference operator for inversion-free trajectory editing. However, face swapping additionally requires a reference-specific identity condition and local control of the residual to avoid overwriting target facial dynamics. FlowFaceSwap addresses these face-specific requirements through test-time identity-token optimization and structure-aware residual modulation.

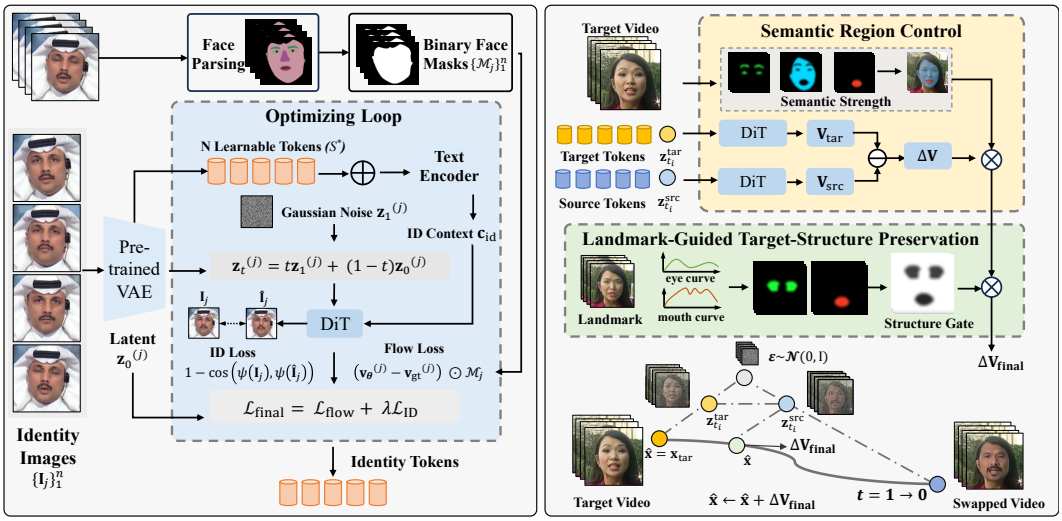
3 METHOD

Given source identity images and a target video, FlowFaceSwap synthesizes a video that transfers the source identity while rigorously preserving the target's inherent attributes, including pose, expressions, and background context. It is built upon a frozen text-to-video rectified flow prior. Rather than training a specialized generator, we first optimize identity tokens at test time to establish a discriminative identity context, followed by an identity-guided flow residual editing process. To prevent structural distortions, we incorporate a structure-aware residual modulation mechanism: temporally tracked semantic masks explicitly control the spatial distribution of the identity residual, while a landmark-guided target-structure preservation strategy dynamically suppresses this residual within expression-critical regions. The overall pipeline is illustrated in Figure 2.

3.1 PRELIMINARIES: RECTIFIED FLOW AND FLOW EDITING

Rectified Flow. Let $\mathbf{x}_0$ denote a clean video latent and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denote a noise with the same spatiotemporal shape. Rectified flow models the linear path

$$\mathbf{z}_t = (1-t)\mathbf{x}_0 + t\boldsymbol{\epsilon}, \quad t \in [0,1], \tag{1}$$



(a) Test-Time Identity Token Optimization *(b) Identity-Guided Flow Residual Editing*

Figure 2: Overview of FlowFaceSwap. We first learn identity tokens at test time with a frozen rectified flow video prior. During generation, velocity predictions conditioned on the source and target identities are compared along the target trajectory to form an identity flow residual. The residual is first localized by temporally tracked semantic regions and is then suppressed in dynamic mouth and eye regions by landmark-guided target-structure preservation.

where ground-truth velocity is $\boldsymbol{\epsilon} - \mathbf{x}_0$. A conditional rectified flow video prior $\mathbf{v}_\theta$ is trained with

$$\mathcal{L}_{\mathrm{RF}} = \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon},t,\mathbf{c}}\left[\left\|\mathbf{v}_\theta(\mathbf{z}_t,t,\mathbf{c}) - (\boldsymbol{\epsilon}-\mathbf{x}_0)\right\|_2^2\right], \tag{2}$$

where $\mathbf{c}$ denotes the conditioning context. With a guidance scale s, the classifier-free guided velocity is defined as

$$\begin{aligned}\mathbf{V}(\mathbf{z},t,\mathbf{c};s) = \mathbf{v}_\theta(\mathbf{z},t,\mathbf{c}_{\mathrm{neg}}) \\ + s\left[\mathbf{v}_\theta(\mathbf{z},t,\mathbf{c}) - \mathbf{v}_\theta(\mathbf{z},t,\mathbf{c}_{\mathrm{neg}})\right],\end{aligned} \tag{3}$$

where $\mathbf{c}_{\mathrm{neg}}$ is the negative or null context. Standard sampling integrates this velocity from the noise endpoint ($t = 1$) toward the clean endpoint ($t = 0$).

Flow Editing. Unlike traditional diffusion models that require computationally heavy inversion to edit real images, rectified flow enables inversion-free editing via direct trajectory alignment (Kulikov et al., 2025). Specifically, given an original video latent $\mathbf{x}^{\mathrm{orig}}$, its exact noisy state at timestep t is analytically determined as $\mathbf{z}_t^{\mathrm{orig}} = (1-t)\mathbf{x}^{\mathrm{orig}} + t\boldsymbol{\epsilon}$. Let $\mathbf{c}_{\mathrm{edit}}$ and $\mathbf{c}_{\mathrm{orig}}$ denote a desired edit context and the original semantic context, respectively. During the reverse editing process, the running edited clean latent estimate $\hat{\mathbf{x}}$ is initialized as $\mathbf{x}^{\mathrm{orig}}$. To evaluate the edited velocity at the correct noise scale, the current estimate is aligned to the original trajectory:

$$\mathbf{z}_t^{\mathrm{edit}} = \hat{\mathbf{x}} + \mathbf{z}_t^{\mathrm{orig}} - \mathbf{x}^{\mathrm{orig}}. \tag{4}$$

A generic semantic flow residual is then defined as the difference between the velocity predicted under the edit context and the original context:

$$\Delta\mathbf{V}_t = \mathbf{V}(\mathbf{z}_t^{\mathrm{edit}},t,\mathbf{c}_{\mathrm{edit}};s_{\mathrm{edit}}) - \mathbf{V}(\mathbf{z}_t^{\mathrm{orig}},t,\mathbf{c}_{\mathrm{orig}};s_{\mathrm{orig}}). \tag{5}$$

The edited latent estimate is progressively accumulated over the reverse-time solver steps (e.g., using the Euler method) via $\hat{\mathbf{x}} \leftarrow \hat{\mathbf{x}} + \Delta t\,\Delta\mathbf{V}_t$. We adapt this inversion-free editing paradigm specifically for video face swapping in the following sections.

3.2 Test-Time Identity Token Optimization

FlowFaceSwap represents each identity by a small set of learnable token embeddings inserted into a fixed text template. During test-time optimization, the video generation model remains frozen and

Algorithm 1 Structure-Preserved Identity Flow Editing

Input: Target video latent $\mathbf{x}^{\mathrm{tar}}$, source context $\mathbf{c}_{\mathrm{src}}$, target context $\mathbf{c}_{\mathrm{tar}}$, negative context $\mathbf{c}_{\mathrm{neg}}$, semantic strength map $\mathcal{A}$, structure regions, timesteps $\{t_i\}_{i=K}^{0}$, guidance scales $s_{\mathrm{src}}, s_{\mathrm{tar}}$.
Output: Swapped video latent $\hat{\mathbf{x}}$.

Initialize $\hat{\mathbf{x}} = \mathbf{x}^{\mathrm{tar}}$
for $i = K$ **down to** 1 **do**
  Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{z}_{t_i}^{\mathrm{tar}} = (1 - t_i)\mathbf{x}^{\mathrm{tar}} + t_i\boldsymbol{\epsilon}$
  $\mathbf{z}_{t_i}^{\mathrm{edit}} = \hat{\mathbf{x}} + \mathbf{z}_{t_i}^{\mathrm{tar}} - \mathbf{x}^{\mathrm{tar}}$
  Compute $\Delta\mathbf{V}_{t_i}$ using Eq. 9
  $\Delta\mathbf{V}_{t_i}^{\mathrm{sem}} = \mathcal{A} \odot \Delta\mathbf{V}_{t_i}$
  Build landmark-guided structure gate $\mathcal{G}_{t_i}$
  $\Delta\mathbf{V}_{t_i}^{\mathrm{final}} = \mathcal{G}_{t_i} \odot \Delta\mathbf{V}_{t_i}^{\mathrm{sem}}$
  $\hat{\mathbf{x}} \leftarrow \hat{\mathbf{x}} + (t_{i-1} - t_i)\Delta\mathbf{V}_{t_i}^{\mathrm{final}}$
end for
return $\hat{\mathbf{x}}$

only these identity tokens are updated. Source tokens are learned from source identity references, while target tokens are learned from target frames to provide the baseline identity context used in Eq. 9.

Given identity frames $\{\mathbf{I}_j\}_{j=1}^{K}$, we encode them into latents $\{\mathbf{x}^{(j)}\}_{j=1}^{K}$ and obtain face masks $\{\mathcal{M}^{(j)}\}_{j=1}^{K}$ from face parsing. Let $\mathbf{c}$ be the context produced by inserting learnable tokens $\mathbf{e}$ into the template. Omitting the identity index j for brevity, the masked rectified flow objective is:

$$\mathcal{L}_{\mathrm{flow}}(\mathbf{e}) = \mathbb{E}_{\mathbf{x},t,\boldsymbol{\epsilon}}\left[w(t)\,\|\mathcal{M} \odot (\mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{c}) - \mathbf{u})\|_2^2\right], \tag{6}$$

where $\mathbf{z}_t = (1 - t)\mathbf{x} + t\boldsymbol{\epsilon}$, $w(t)$ is a timestep weight, and $\mathbf{u} = \boldsymbol{\epsilon} - \mathbf{x}$ denotes the ground-truth flow velocity. The mask restricts the reconstruction signal to the facial region, reducing the tendency of the token to absorb background or frame-specific content.

The reconstruction objective alone may produce visually plausible faces with ambiguous identity. We therefore introduce an identity loss with a frozen ArcFace recognizer (Deng et al., 2019). Let $\psi(\cdot)$ denote the recognizer and let $\hat{\mathbf{I}}^{(j)}(\mathbf{e})$ be the clean image estimated with the current token. We define the identity loss as:

$$\mathcal{L}_{\mathrm{ID}}(\mathbf{e}) = 1 - \cos\left(\psi(\mathrm{crop}(\hat{\mathbf{I}}^{(j)}(\mathbf{e}))), \psi(\mathrm{crop}(\mathbf{I}_{\mathrm{ref}}))\right), \tag{7}$$

where $\mathbf{I}_{\mathrm{ref}}$ is a reference image of the same identity. Combining Eq. 6 and Eq. 7, the final optimization objective is:

$$\mathcal{L}_{\mathrm{final}} = \mathcal{L}_{\mathrm{flow}} + \lambda_{\mathrm{ID}}\mathcal{L}_{\mathrm{ID}}. \tag{8}$$

This test-time optimization improves identity discriminability without updating the generator or introducing a trained face-swapping module.

3.3 Identity-Guided Flow Residual Editing

Building upon the generic editing formulation (Eq. 5), we formalize video face swapping by treating the target video as the original state ($\mathbf{x}^{\mathrm{orig}} = \mathbf{x}^{\mathrm{tar}}$). The edit and original contexts are instantiated using the optimized source and target identity tokens. The identity-guided flow residual is thus computed as:

$$\Delta\mathbf{V}_t = \mathbf{V}(\mathbf{z}_t^{\mathrm{edit}}, t, \mathbf{c}_{\mathrm{src}}; s_{\mathrm{src}}) - \mathbf{V}(\mathbf{z}_t^{\mathrm{tar}}, t, \mathbf{c}_{\mathrm{tar}}; s_{\mathrm{tar}}). \tag{9}$$

Applying this identity residual uniformly over the target latent would make identity transfer strong but can also distort target articulation, eye motion, and local temporal structure. To mitigate these distortions, we view $\Delta\mathbf{V}_t$ as a raw identity signal that should be spatially and temporally regulated through a two-stage modulation process. In the first stage, the semantic strength map localizes the identity residual to face regions with different strengths; in the second stage, the structure gate then

suppresses these localized residuals in mouth and eye regions where target motion should dominate. Formally, let $\mathcal{A}$ denote the semantic region strength map and $\mathcal{G}_t$ denote the landmark-guided structure gate. The modulated residual is thus computed in two stages:

$$\begin{cases} \Delta \mathbf{V}_t^{\text{sem}} = \mathcal{A} \odot \Delta \mathbf{V}_t, \\ \Delta \mathbf{V}_t^{\text{final}} = \mathcal{G}_t \odot \Delta \mathbf{V}_t^{\text{sem}}. \end{cases} \tag{10}$$

Finally, the edited latent is updated as $\hat{\mathbf{x}} \leftarrow \hat{\mathbf{x}} + \Delta t \, \Delta \mathbf{V}_t^{\text{final}}$, where $\Delta t = t_{i-1} - t_i$ denotes the reverse-time solver step.

3.4 Semantic Region Control

We now elaborate on the construction of the semantic strength map $\mathcal{A}$ in Eq. 10. The key insight is that facial regions contribute unequally to identity preservation and expression representation. Motivated by these regional differences, we propose to derive temporally tracked semantic masks from the target video to spatially modulate the identity residual.

For each target frame, face parsing produces soft masks for the full face, mouth, and eyes/brows. These masks are resized to the latent video grid and combined into four regions: $\{\mathcal{S}_{\text{id}}, \mathcal{S}_{\text{mouth}}, \mathcal{S}_{\text{eye}}, \mathcal{S}_{\text{bd}}\}$, corresponding to identity core, mouth, eyes/brows, and facial boundary. The identity core is obtained by excluding mouth, eyes/brows, and boundary regions from the full face mask. To ensure temporal coherence, we apply lightweight temporal smoothing only to the stable identity mask; dynamic masks remain per-frame to faithfully track the target's instantaneous facial movements. Finally, we aggregate the isolated region masks into a unified spatial weight map, which is defined as a weighted composite: $\mathcal{A} = \sum_{r \in \mathcal{R}} \gamma_r \mathcal{S}_r, \quad \mathcal{R} = \{\text{id}, \text{mouth}, \text{eye}, \text{bd}\}$. The coefficients $\{\gamma_r\}$ encode region-specific modulation strategies: stable identity regions receive strong residuals, while the expression region is assigned a reduced residual strength to better preserve target speech articulation.

3.5 Landmark-Guided Target-Structure Preservation

While the semantic strength map $\mathcal{A}$ effectively localizes edits, its regional weights remain static across all frames. In fact, expression intensity varies significantly over time. For example, a wide-open mouth during speech requires much stronger structural protection than a neutrally closed mouth. To address this temporal variation, we introduce a landmark-guided structure gate $\mathcal{G}_t$ that dynamically adapts the residual suppression according to instantaneous target movements.

For each target frame f, facial landmarks provide soft latent masks for the outer lips, inner mouth, and eyes. Let $\mathcal{L}_{\text{mouth}}$ denote the combined mask of the outer lips and inner mouth, and $\mathcal{L}_{\text{eye}}$ denote the mask of the eyes. We also compute target-structure curves: mouth aperture a_f, eye closure b_f, and their temporal changes. The normalized mouth and eye dynamic scores are:

$$\begin{cases} \tilde{d}_f^{\text{mouth}} = \eta a_f + (1-\eta)|a_f - a_{f-1}|, \\ \tilde{d}_f^{\text{eye}} = \eta b_f + (1-\eta)|b_f - b_{f-1}|, \end{cases} \tag{11}$$

where η balances the instantaneous structure value and its temporal variation. These raw scores are then min-max normalized across the video sequence to $[0, 1]$, yielding the final dynamic scores d_f^{mouth} and d_f^{eye}.

The structure gate is also timestep dependent. Let $\ell(t) = \max\left(1 - \frac{t}{\tau}, 0\right)$ be a late-stage weight, where $t = 0$ denotes the clean endpoint. For a specific structural region $r \in \{\text{mouth}, \text{eye}\}$, the gate value is defined as

$$g_{t,f}^r = \max\left(1 - \ell(t)(1-\rho_r) - \beta d_f^r, \rho_r\right). \tag{12}$$

The lower bound ρ_r avoids eliminating the residual entirely, while β controls dynamic suppression. The final spatiotemporal gate integrates these regional modulations:

$$\mathcal{G}_t = \prod_{r \in \{\text{mouth}, \text{eye}\}} \left[1 - \mathcal{L}_r(1 - g_{t,f}^r)\right]. \tag{13}$$

Method	FF++						CelebVHQ					
	FVD ↓	Sm. ↑	Top-1 ↑	Top-5 ↑	Expr ↓	Pose ↓	FVD ↓	Sm. ↑	Top-1 ↑	Top-5 ↑	Expr ↓	Pose ↓
Training-based												
FSGAN	417.83	0.9915	0.54	0.86	0.4834	4.37	125.58	0.9730	0.29	0.55	0.3678	8.61
CanonSwap	230.49	0.9912	0.94	0.96	0.5101	4.20	112.61	0.9739	0.85	0.97	0.4855	7.80
VividFace	238.30	0.9906	0.92	0.94	0.5041	4.44	146.94	0.9737	0.42	0.63	0.5007	7.96
DreamID-V	360.00	0.9929	0.96	0.96	0.5245	4.32	221.64	0.9739	0.99	1.00	0.5373	7.74
Stand-In	665.75	0.9912	0.82	0.90	0.4935	5.67	789.24	0.9751	0.53	0.76	0.5756	10.05
Methods without face-swap training												
PhotoSwap	773.49	0.9784	0.74	**0.98**	**0.4616**	4.64	682.23	0.9521	0.29	0.94	**0.4674**	12.84
SwapAnything	825.57	0.9857	0.74	0.96	0.4776	5.03	557.74	0.9669	0.57	0.92	0.4950	9.42
FlowDirector	366.57	0.9893	**0.92**	**0.98**	0.5154	4.68	677.92	**0.9916**	0.74	0.94	0.5440	9.58
FlowFaceSwap	**313.02**	**0.9901**	**0.92**	**0.98**	0.4939	**4.53**	**250.47**	0.9719	**0.78**	**0.96**	0.5237	8.68
Methods using an external image face-swapper												
VFace	338.89	0.9896	0.96	0.98	0.4949	4.37	174.16	0.9719	0.79	0.98	0.5071	7.85
AnyV2V	425.06	0.9849	0.88	0.96	0.4998	4.64	369.90	0.9689	0.45	0.86	0.4880	8.52

Table 1: Main quantitative comparison on FF++ and CelebVHQ.

Outside the structural mask regions $\mathcal{L}_r$, $\mathcal{G}_t$ remains close to one. In dynamic expression frames and near the clean endpoint, the gate more strongly suppresses the residual within $\mathcal{L}_r$, safeguarding the target structure during complex dynamic motions.

4 Experiments

4.1 Experimental Protocol

Datasets We evaluate FlowFaceSwap on two widely used video face-swapping benchmarks. For FaceForensics++ (FF++) (Rössler et al., 2019), we utilize 50 source-target identity pairs and extract the initial 61 frames of each target video as the driving sequence. To assess robustness on higher-quality, in-the-wild portraits, we curate an additional 100 source-target pairs from CelebVHQ (Zhu et al., 2022), maintaining the identical 61-frame configuration.

Metrics Following standard protocols, we evaluate video quality, temporal consistency, identity fidelity, and attribute preservation. Video quality is assessed using the Fréchet Video Distance (FVD, ↓) (Unterthiner et al., 2019). Furthermore, we report Smoothness (↑) (Huang et al., 2024) to evaluate temporal consistency. Identity fidelity is quantified via Top-1 and Top-5 overall ID retrieval accuracy (↑) using an ArcFace recognizer (Deng et al., 2019). Attribute preservation is evaluated using expression error (↓), measured by the L2 distance between the expression features of the generated and target frames, and pose error (↓), derived from estimated Euler angles. To prevent evaluation bias, the identity loss and identity evaluation employ different frozen ArcFace recognizers.

Baselines Our baselines are grouped into three categories. The first category consists of training-based methods, including FSGAN (Nirkin et al., 2019), CanonSwap (Luo et al., 2025), VividFace (Shao et al., 2025), DreamID-V (Guo et al., 2026), and Stand-In (Xue et al., 2026). The second category contains methods without benchmark-specific face-swap training, namely PhotoSwap (Gu et al., 2023), SwapAnything (Gu et al., 2025), and FlowDirector (Li et al., 2025). Since FlowDirector is a generic video-editing method without native face-swapping support, we adapt its direct residual formulation using the same test-time source and target identity-token optimization as FlowFaceSwap, while omitting the proposed semantic and landmark controls. This category provides the closest comparison under the training-free setting. The third category includes VFace (Baliah et al., 2026) and AnyV2V (Ku et al., 2024), which rely on an external image-based face-swapper in their pipelines and are therefore reported separately.

Implementation Details We use the frozen Wan2.1-T2V-1.3B model (Wan et al., 2025) to generate 512×512 videos with 61 frames. We optimize one identity token for each source and target identity for 2,000 steps with learning rate 10^{-3}, using a frozen ArcFace identity loss with $\lambda_{\mathrm{ID}} = 0.001$; flow editing uses 50 sampling steps with $(s_{\mathrm{src}}, s_{\mathrm{tar}}, \mathrm{shift}) = (10.5, 3.5, 5.0)$. All experiments run



Figure 3: Qualitative comparison with training-free baselines on FF++ and CelebVHQ.

on a single 24GB NVIDIA RTX 3090 GPU; further implementation details are provided in Appendix A.1.

4.2 Comparison with Baselines

Quantitative Evaluation. Table 1 summarizes the quantitative results on FF++ and CelebVHQ. Among methods without face-swap training, PhotoSwap and SwapAnything achieve relatively low expression errors on FF++ (0.4616 and 0.4776), but provide limited identity retrieval (Top-1 0.74) and high FVD scores (773.49 and 825.57). FlowDirector improves identity retrieval to 0.92, but its higher expression error and FVD indicate a less favorable balance between identity transfer and target-video fidelity. FlowFaceSwap achieves a stronger overall balance within this group. On FF++, it obtains the lowest FVD (313.02), lowest pose error (4.53), and highest temporal smoothness (0.9901), while retaining a Top-1 retrieval accuracy of 0.92. On CelebVHQ, it achieves an FVD of 250.47 and a Top-1 retrieval accuracy of 0.78, maintaining this balance under more challenging in-the-wild conditions. For methods using an external image swapper, VFace achieves higher identity retrieval than FlowFaceSwap, while our method obtains lower FVD on FF++. AnyV2V has lower identity retrieval and higher FVD than FlowFaceSwap on both datasets. Despite competitive results on selected metrics, their dependence on an external face swapper constitutes both a practical limitation and a distinct inference setting relative to FlowFaceSwap.

Qualitative Evaluation. As shown in Figure 3, existing training-free baselines struggle to preserve target structures during identity injection. PhotoSwap exhibits poor identity consistency, while SwapAnything often introduces facial distortions. Although AnyV2V retains identity slightly better, it yields an unnatural appearance. FlowDirector distorts complex expressions, leading to noticeable intra-oral artifacts and structural leakage from source identity. In contrast, FlowFaceSwap executes precise geometric alignment and clean semantic blending. It preserves target dynamics without sacrificing the source identity, demonstrating superior fidelity.

Runtime and Memory. We measure runtime and peak GPU memory under the same hardware setting. As shown in Table 3, FlowFaceSwap requires 13.30 sec/frame with a peak GPU memory usage of 21.8 GB, substantially lower than SwapAnything (41.80 sec/frame) and AnyV2V (20.05 sec/frame). This lower swap-stage latency is consistent with our inversion-free trajectory editing, whereas SwapAnything and AnyV2V involve additional inversion-based processing. Notably, FlowFaceSwap requires a one-time optimization of the identity token for each identity, which takes approximately 33 minutes per identity on a single NVIDIA RTX 3090 GPU. Once optimized, the token

Variant	FVD ↓	Top-1 ↑	Top-5 ↑	Expr ↓	Pose ↓
w/o ID loss	325.41	0.86	0.96	0.5107	4.51
w/o sem. map	329.17	0.94	0.98	0.4952	4.49
w/o struct. gate	347.16	0.94	0.98	0.5098	4.59
Full	313.02	0.92	0.98	0.4939	4.53

Table 2: Core component ablation on FF++.

Method	Params	Sec.	Mem.
SwapAnything	0.9B	41.80	19.8 GB
AnyV2V	1.4B	20.05	23.3 GB
FlowFaceSwap	1.3B	13.30	21.8 GB

Table 3: Runtime and GPU memory.

is cached and reused in subsequent face-swapping runs. Thus, the cost is incurred once per identity rather than for every video frame.

4.3 Ablation Study

We conduct ablation studies on the FF++ dataset to evaluate the contribution of each core component and our spatial modulation strategies. Additional ablation studies and qualitative visualizations are provided in the Appendix.

Efficacy of Core Components. The component ablation in Table 2 evaluates the contribution of identity loss, semantic region control, and landmark-guided structure preservation. Removing the identity loss lowers Top-1 retrieval from 0.92 to 0.86, suggesting that flow reconstruction alone provides insufficient identity discrimination. Replacing the precise semantic map with a global mask, or removing the landmark-guided structure gate, degrades video quality and increases expression error. The FlowDirector row in Table 1 offers a complementary reference based on an unmodulated residual. It achieves comparable identity retrieval, but produces lower video quality and weaker structure preservation. Overall, identity learning strengthens the source representation, while semantic and landmark controls prevent the residual from overwriting fine target motions.

Region	Strength	FVD ↓	Top-1 ↑	Top-5 ↑	Expr ↓	Pose ↓
Identity region						
Identity	0.85	254.67	0.76	0.96	0.4818	4.51
Identity	1.00	313.02	**0.92**	**0.98**	0.4939	**4.53**
Mouth region						
Mouth	0.85	**313.02**	**0.92**	**0.98**	0.4939	**4.53**
Mouth	1.00	318.81	0.92	0.98	0.4933	4.58
Eye/brow region						
Eye/brow	0.85	313.02	**0.92**	**0.98**	**0.4939**	**4.53**
Eye/brow	1.00	312.30	0.90	0.96	0.4955	4.56
Boundary region						
Boundary	0.70	**313.02**	**0.92**	**0.98**	**0.4939**	**4.53**
Boundary	1.00	317.16	0.92	0.98	0.4958	4.61

Table 4: Semantic strength map analysis on FF++.

Precise Semantic Modulation. Table 4 demonstrates how the semantic map affords granular spatial control. Preserving the full residual strength (1.00) in stable identity regions is imperative for high retrieval accuracy; reducing it to 0.85 improves FVD (254.67) but severely compromises identity transfer (Top-1 drops to 0.76). However, applying this strength uniformly to expression-sensitive zones, such as the mouth or eyes, unexpectedly inflates the FVD and disrupts natural articulation. By selectively attenuating the flow residual in these localized regions, FlowFaceSwap confines identity transfer within target areas while avoiding global structural degradation.

5 Conclusion

We present FlowFaceSwap, showing that high-fidelity video face swapping is achievable without task-specific training. By combining test-time identity optimization with inversion-free flow residual editing, our framework overcomes the computational challenges of conventional diffusion pipelines. We demonstrate that explicit spatial regulation via temporally tracked semantics and landmark-guided gates is indispensable for decoupling identity from target dynamics. This approach sets a new standard among training-free methods, delivering generation quality and structural preservation on par with specialized models. Future work will explore integrating deeper 3D geometric priors to tackle challenges under extreme occlusions and profile poses.

AI USE STATEMENT

In this work, we used generative AI tools to help locate relevant literature and to polish the wording and readability of the manuscript. We have not used generative AI tools to generate synthetic data, develop the method or theoretical framework, formulate or prove mathematical claims, design or execute experiments, translate the manuscript, clean or reformat datasets, or interpret experimental results. We reviewed all AI-assisted work, manually verified the cited literature and final text, and take responsibility for the final content of this work.

ETHICS STATEMENT

Video face swapping can support applications such as entertainment and privacy-preserving anonymization, but it can also be misused to create non-consensual imagery, impersonation, and misinformation. This work is intended for academic research on flow-based identity transfer. No human subjects were recruited and no new identity dataset was collected. We use the FF++ and CelebVHQ benchmarks under their respective access and research-use terms, and do not release a new dataset containing real identities. The code will be released for academic research under a research-only license.

REPRODUCIBILITY STATEMENT

This paper presents the complete formulation of FlowFaceSwap. The experimental section specifies the datasets, the source–target pairing protocol, the evaluation metrics, the baselines, and the inference settings. Appendix provides additional implementation details. The supplementary material includes the core implementation and inference configurations of FlowFaceSwap. Upon acceptance, we will also publicly release the source code.

REFERENCES

Sanoojan Baliah, Qinliang Lin, Shengcai Liao, Xiaodan Liang, and Muhammad Haris Khan. Realistic and efficient face swapping: A unified approach with diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1062–1071. IEEE, 2025.

Sanoojan Baliah, Yohan Abeysinghe, Rusiru Thushara, Khan Muhammad, Abhinav Dhall, Karthik Nandakumar, and Muhammad Haris Khan. Vface: A training-free approach for diffusion-based video face swapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4315–4324, 2026.

Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K. Nayar. Face swapping: automatically replacing faces in photographs. In *ACM SIGGRAPH 2008 Papers*, SIGGRAPH '08, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781450301121. doi: 10.1145/1399504.1360638. URL https://doi.org/10.1145/1399504.1360638.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, pp. 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co. ISBN 0201485605. doi: 10.1145/311535.311556. URL https://doi.org/10.1145/311535.311556.

Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging faces in images. *Computer Graphics Forum*, 23(3):669–676, August 2004. ISSN 1467-8659. doi: 10.1111/j.1467-8659.2004.00799.x. URL http://dx.doi.org/10.1111/j.1467-8659.2004.00799.x.

Weston Bondurant, Arkaprava Sinha, Hieu Le, Srijan Das, and Stephanie Schuckers. Diffswap++: 3d latent-controlled diffusion for identity-preserving face swapping. *arXiv preprint arXiv:2511.05575*, 2025.

Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on*

Multimedia, MM '20, pp. 2003–2011, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413630. URL https://doi.org/10.1145/3394171.3413630.

Yiyang Chen, Xuanhua He, Xiujun Ma, and Jack Ma. Contextflow: Training-free video object editing via adaptive context enrichment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(4):3129–3137, March 2026. ISSN 2159-5399. doi: 10.1609/aaai.v40i4.37306. URL http://dx.doi.org/10.1609/aaai.v40i4.37306.

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.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

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.

Jing Gu, Yilin Wang, Nanxuan Zhao, Tsu-Jui Fu, Wei Xiong, Qing Liu, Zhifei Zhang, HE Zhang, Jianming Zhang, HyunJoon Jung, and Xin Eric Wang. Photoswap: Personalized subject swapping in images. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 35202–35217. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6e9a0a72da9b76c3ebc8cc33ff10ac29-Paper-Conference.pdf.

Jing Gu, Nanxuan Zhao, Wei Xiong, Qing Liu, Zhifei Zhang, He Zhang, Jianming Zhang, HyunJoon Jung, Yilin Wang, and Xin Eric Wang. Swapanything: Enabling arbitrary object swapping in personalized image editing. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 402–418, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73411-3.

Xu Guo, Fulong Ye, Xinghui Li, Pengqi Tu, Pengze Zhang, Qichao Sun, Songtao Zhao, Xiangwang Hou, and Qian He. Dreamid-v: Bridging the image-to-video gap for high-fidelity face swapping via diffusion transformer. *arXiv preprint arXiv:2601.01425*, 2026.

Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. *Advances in neural information processing systems*, 37:36777–36804, 2024.

Yue Han, Junwei Zhu, Keke He, Xu Chen, Yanhao Ge, Wei Li, Xiangtai Li, Jiangning Zhang, Chengjie Wang, and Yong Liu. Face-adapter for pre-trained diffusion models with fine-grained id and attribute control. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part L*, pp. 20–36, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72972-0. doi: 10.1007/978-3-031-72973-7_2. URL https://doi.org/10.1007/978-3-031-72973-7_2.

Dailan He, Xiahong Wang, Shulun Wang, Guanglu Song, Bingqi Ma, Hao Shao, Yu Liu, and Hongsheng Li. High-fidelity diffusion face swapping with id-constrained facial conditioning. *arXiv preprint arXiv:2503.22179*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Taewoong Kang, Sohyun Jeong, Hyojin Jang, and Jaegul Choo. Zero-shot head swapping in real-world scenarios. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10805–10814, 2025. doi: 10.1109/CVPR52734.2025.01009.

Jeongsol Kim, Yeobin Hong, Jonghyun Park, and Jong Chul Ye. Flowalign: Trajectory-regularized, inversion-free flow-based image editing, 2025a. URL https://arxiv.org/abs/2505.23145.

Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. Diffface: Diffusion-based face swapping with facial guidance. *Pattern Recognition*, 163:111451, 2025b. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2025.111451. URL https://www.sciencedirect.com/science/article/pii/S0031320325001116.

Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=RFrJCkw2oa. Reproducibility Certification.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19721–19730, October 2025.

Jaeseong Lee, Junha Hyung, Sohyun Jung, and Jaegul Choo. Selfswapper: Self-supervised face swapping via shape agnostic masked autoencoder. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LV*, pp. 383–400, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73000-9. doi: 10.1007/978-3-031-73001-6_22. URL https://doi.org/10.1007/978-3-031-73001-6_22.

Guangzhao Li, Yanming Yang, Chenxi Song, and Chi Zhang. Flowdirector: Training-free flow steering for precise text-to-video editing, 2025. URL https://arxiv.org/abs/2506.05046.

Tianle Li, Max Ku, Cong Wei, and Wenhu Chen. Dreamedit: Subject-driven image editing. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=P9haooN9v2.

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. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=XVjTT1nw5z.

Zhian Liu, Maomao Li, Yong Zhang, Cairong Wang, Qi Zhang, Jue Wang, and Yongwei Nie. Fine-grained face swapping via regional gan inversion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8578–8587, 2023b. doi: 10.1109/CVPR52729.2023.00829.

Xiangyang Luo, Ye Zhu, Yunfei Liu, Lijian Lin, Cong Wan, Zijian Cai, Yu Li, and Shao-Lun Huang. Canonswap: High-fidelity and consistent video face swapping via canonical space modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10064–10074, October 2025.

Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7184–7193, 2019.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510, 2023. doi: 10.1109/CVPR52729.2023.02155.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–11, 2019. doi: 10.1109/ICCV.2019.00009.

Hao Shao, Shulun Wang, Yang Zhou, Guanglu Song, Dailan He, Zhuofan Zong, Shuo Qin, Yu Liu, and Hongsheng Li. Vividface: A robost and high-fidelity video face swapping framework. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=wyv81ezGgv.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8543–8552, 2024. doi: 10.1109/CVPR52733.2024.00816.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2019. URL https://arxiv.org/abs/1812.01717.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

Runqi Wang, Yang Chen, Sijie Xu, Tianyao He, Wei Zhu, Dejia Song, Nemo Chen, Xu Tang, and Yao Hu. Dynamicface: High-quality and consistent face swapping for image and video using composable 3d facial priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13438–13447, October 2025.

Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping, 2021. URL https://arxiv.org/abs/2106.09965.

Zhen Wang, Youcan Xu, Jun Xiao, and Long Chen. Flowmotion: Training-free flow guidance for video motion transfer. *arXiv preprint arXiv:2603.06289*, 2026.

Bowen Xue, Zheng-Peng Duan, Qixin Yan, Wenjing Wang, Hao Liu, Chun-Le Guo, Chongyi Li, Chen Li, and Jing Lyu. Stand-in: A lightweight and plug-and-play identity control for video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23314–23324, 2026.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18381–18391, 2023. doi: 10.1109/CVPR52729.2023.01763.

Kaixiang Yang, Boyang Shen, Xin Li, Yuchen Dai, Yuxuan Luo, Yueran Ma, Wei Fang, Qiang Li, and Zhiwei Wang. Fia-edit: Frequency-interactive attention for efficient and high-fidelity inversion-free text-guided image editing, 2025. URL https://arxiv.org/abs/2511.12151.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Daehyun Yoo, Hongchul Lee, and Jiho Kim. Inversion-based face swapping with diffusion model. *IEEE Access*, 13:6764–6774, 2025. doi: 10.1109/ACCESS.2024.3519163.

Zheng Yu, Yaohua Wang, Siying Cui, Aixi Zhang, Wei-Long Zheng, and Senzhang Wang. Fuseanypart: Diffusion-driven facial parts swapping via multiple reference images. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 80864–80884. Curran Associates, Inc., 2024. doi: 10.52202/079017-2569. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/939f20cc178460749e4ab5fa28fd3b10-Paper-Conference.pdf.

Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8568–8577, 2023. doi: 10.1109/CVPR52729.2023.00828.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pp. 650–667. Springer, 2022.

Yuhao Zhu, Qi Li, Jian Wang, Chengzhong Xu, and Zhenan Sun. One shot face swapping on megapixels. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4832–4842, 2021. doi: 10.1109/CVPR46437.2021.00480.

A Appendix

A.1 Experimental Details

All results use the frozen Wan2.1-T2V-1.3B generator at 512-pixel resolution. Unless stated otherwise, we optimize one identity token for 2000 steps with Adam at a learning rate of 10^{-3} and set the ArcFace identity-loss weight to $\lambda_{\mathrm{ID}} = 0.001$. Sampling uses 50 steps, a shift of 5.0, and guidance scales of $s_{\mathrm{src}} = 10.5$ for the source-ID editing branch and $s_{\mathrm{tar}} = 3.5$ for the target-identity anchor branch. The region weights are $\{\gamma_{\mathrm{id}}, \gamma_{\mathrm{mouth}}, \gamma_{\mathrm{eye}}, \gamma_{\mathrm{bd}}\} = \{1.0, 0.85, 0.85, 0.70\}$. For landmark-guided preservation, the minimum residual strengths for mouth and eye regions are 0.30 and 0.35, the dynamic coefficient is $\beta = 0.50$, and the late-stage ratio is $\tau = 0.25$. No generator parameters are updated.

For each target frame, a RetinaFace-Mobilenet detector and the FaRL-LaPa-448 parser provide face, mouth, and eye/brow masks. These masks are projected to the 64×64 spatial latent grid. The identity region excludes the mouth, eye/brow, and boundary regions from the face support: $\mathcal{S}_{\mathrm{id}} = \mathcal{S}_{\mathrm{face}} \setminus (\mathcal{S}_{\mathrm{mouth}} \cup \mathcal{S}_{\mathrm{eye}} \cup \mathcal{S}_{\mathrm{bd}})$. The boundary is defined as the ring between the face mask and its one-pixel 3×3 erosion at the latent-mask resolution. To reduce segmentation jitter without blurring articulation, a replicate-padded five-frame average is applied to the identity, boundary, and full-face masks only; mouth and eye masks remain frame-adaptive, and a 3×3 spatial average is then applied to all region masks. Landmark trajectories are obtained with MediaPipe FaceMesh using 0.5 detection and tracking confidences; when unavailable, we use a 68-point face-alignment fallback. For an isolated parsing failure after a valid frame, the preceding valid semantic masks and structure prior are propagated.

All baselines are evaluated with their official implementations and default inference settings on the same source-target pairs, target-frame windows, and dataset splits. Token optimization and identity evaluation use independently frozen recognizers. For retrieval, we scan up to 30 source frames per identity, rank valid detections by detection confidence, face area, and Laplacian sharpness, and average the top 10 embeddings; at least three valid source frames are required. The gallery consists of all source identities in the evaluation set (50 for FF++ and 100 for CelebVHQ), each represented by the averaged embedding of its top valid frames. Each generated video is represented by the average embedding of its first 20 valid frames. We apply no post-hoc detection-score or face-area filtering, so difficult frames are not removed from the protocol; frames without a valid identity feature are excluded from embedding aggregation, and a video without any valid embedding is counted as a retrieval failure. Expression error is the L2 distance between pretrained FECNet features extracted from 224×224 face crops. Pose is estimated from 3D facial landmarks through a perspective-n-point fit, with failed estimates assigned a 90-degree penalty in the all-frame aggregate. FVD is computed

Method	Hard Mouth		Hard Eye	
	Top-1↑	Expr↓	Top-1↑	Expr↓
FSGAN	0.52	0.4841	0.60	0.4910
CanonSwap	1.00	0.5783	1.00	0.5584
VividFace	0.92	0.5160	0.92	0.4961
DreamID-V	1.00	0.6265	1.00	0.6139
Stand-In	0.85	0.5644	0.85	0.5673
PhotoSwap	0.62	0.4974	0.62	0.5152
SwapAnything	0.69	0.5124	0.69	0.4928
FlowDirector	0.92	0.5400	0.92	0.5526
AnyV2V	0.92	0.6074	0.92	0.5979
FlowFaceSwap (Ours)	**1.00**	0.5163	**1.00**	0.5176

Table 5: Performance on hard dynamic-motion subsets of FF++. Hard Mouth and Hard Eye are the top 25% source-target pairs ranked by target mouth and eye landmark acceleration, respectively ($n = 13$ for each subset).



Figure 4: Visualization of the rectified flow editing trajectory. The first two columns show the source reference and target frame; subsequent columns show decoded edited-latent states saved from the actual sampling process at the indicated steps.

from target and swapped clips resized to 224×224, and Smoothness is measured by the VBench AMT evaluator.

A.2 Dynamic Mouth/Eye Preservation

The main paper reports aggregate expression and pose errors over all frames. To evaluate target-structure preservation under more difficult motion, we further evaluate hard FF++ subsets selected by target landmark dynamics. Specifically, videos are ranked by target mouth landmark acceleration and target eye landmark acceleration, and the top 25% pairs are used as hard talking and blinking subsets. Table 5 shows that the hard subsets expose a sharper identity-motion tradeoff than the aggregate FF++ table. Training-free methods such as PhotoSwap and SwapAnything preserve some target expressions, but their Top-1 identity retrieval drops to 0.62–0.69, indicating that they often remain close to the target identity rather than completing the swap. FlowDirector and AnyV2V retain stronger identity, but their expression errors are substantially higher than ours on both hard subsets. Training-based methods are competitive in the main quantitative table, yet their behavior is less consistent under these challenging target motions: FSGAN, VividFace, and Stand-In lose identity, while CanonSwap and DreamID-V keep identity at the cost of much larger expression error. This suggests that specialized face-swap training does not fully cover difficult mouth and eye dynamics, limiting their generalization to challenging and unseen scenarios. Among methods with perfect Top-1 retrieval on both hard subsets, FlowFaceSwap achieves the lowest expression error, outperforming CanonSwap and DreamID-V by 0.0620/0.1102 on Hard Mouth and 0.0408/0.0963 on Hard Eye. These results confirm that our structure-preserved flow editing achieves precise temporal alignment of mouth and eye motions while injecting the source identity.



Figure 5: Failure cases on CelebVHQ.

A.3 Rectified Flow Editing Trajectory

As illustrated in Figure 4, the rectified flow formulation enables a direct editing path from the target-conditioned generation process. The intermediate columns are decoded edited-latent estimates recorded at selected sampling steps. Source-identity cues become progressively stronger across the saved states, while the target framing and pose remain aligned. This indicates that FlowFaceSwap does not reconstruct a target video through a separate inversion stage and then edit it; it directly accumulates identity residuals during target-conditioned rectified flow sampling. By operating directly on the target video and leveraging identity-token guidance, it naturally preserves the target's expression and pose while replacing the identity.

A.4 Failure Cases

Figure 5 shows two challenging CelebVHQ examples. In the first case, extreme source appearance and makeup are transferred together with identity cues, creating a conspicuous appearance mismatch. The second case involves a large pose gap where the source face is largely unobservable, making it difficult to preserve sufficient identity information and leading to visible artifacts in the swapped result. Collectively, these failure cases indicate that training-free swapping remains fragile under large pose gaps and extreme source appearances.

A.5 Identity Loss Weight and Number of Identity Tokens

Identity loss weight. We ablate the ArcFace-clean identity loss weight used during test-time identity-token optimization. Increasing λ_{ID} strengthens source-identity transfer, but overly strong identity supervision can reduce target expression fidelity and perceptual video quality.

Table 6 shows that the ArcFace identity loss is useful. Without this loss, the optimized token still contains identity information from reconstruction, but Top-1 retrieval remains lower and expression error is higher. Increasing λ_{ID} to 10^{-3} improves Top-1 retrieval to 0.92 while reducing expression

λ_{ID}	Top-1↑	Top-5↑	FVD↓	Smooth↑	Expr↓	Pose↓
0	0.86	0.96	325.41	0.9901	0.5107	4.51
5×10^{-4}	0.86	0.98	318.61	0.9901	0.4992	4.53
10^{-3}	0.92	0.98	313.02	0.9901	0.4939	4.53
3×10^{-3}	0.92	0.98	361.92	0.9901	0.5126	4.53

Table 6: Ablation study on the identity loss weight.

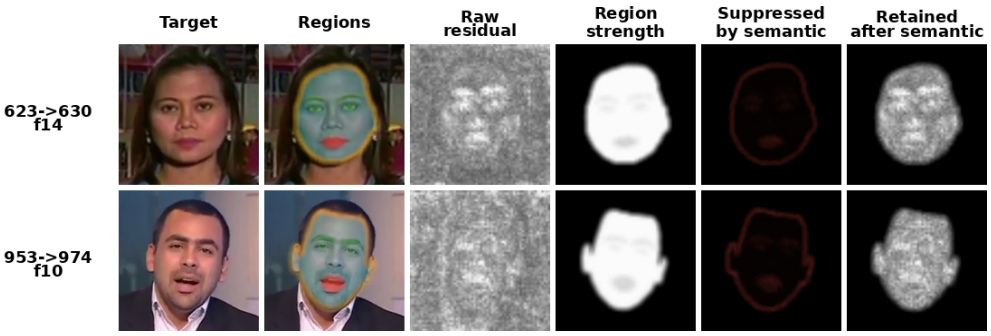


Figure 6: Semantic region control. Columns show the target frame, semantic regions, raw identity-residual magnitude, final region-strength map, residual suppressed by semantic control, and residual retained after semantic control.

error, indicating that discriminative identity optimization helps define a cleaner source direction rather than merely amplifying artifacts. Further increasing the weight to 3×10^{-3} gives no Top-1 improvement, while FVD rises sharply from 313.02 to 361.92 and expression error increases to 0.5126. This confirms that overly strong identity supervision can dominate the target-driven dynamics.

Number of identity tokens. We evaluate whether increasing the number of learnable identity tokens improves source-identity transfer or instead over-parameterizes the test-time identity representation. As shown in Table 7, more identity tokens mainly increase the capacity of the test-time identity representation. Moving from one token to four tokens raises Top-1 retrieval from 0.92 to 0.98. However, FVD increases by more than 100 points, expression error rises from 0.4939 to 0.5278, pose error also increases, and smoothness slightly drops. Since Top-5 retrieval is already saturated with a single token, the additional tokens mostly strengthen identity at the expense of target-motion fidelity.

A.6 Sampling Schedule and Guidance Ablation

We analyze the two sampler-side controls that determine how strongly source-identity residuals act during generation: the sampling shift and the classifier-free guidance scales.

The sampling shift η is the flow-matching noise-schedule parameter used by the Wan sampler. It transforms the nominal noise level σ into $\sigma' = \eta\sigma/(1 + (\eta - 1)\sigma)$. A smaller shift concentrates solver steps near the low-noise endpoint, which improves expression and pose preservation, but leaves less room for identity residuals to reshape the generated face. A larger shift allocates more computation to higher-noise states, where identity residuals have a broader effect on the generated face.

ID tokens	Top-1↑	Top-5↑	FVD↓	Smooth↑	Expr↓	Pose↓
1	0.92	0.98	313.02	0.9901	0.4939	4.53
2	0.94	0.98	337.38	0.9900	0.5245	4.60
4	0.98	0.98	417.33	0.9894	0.5278	4.63

Table 7: Ablation study on the number of identity tokens.

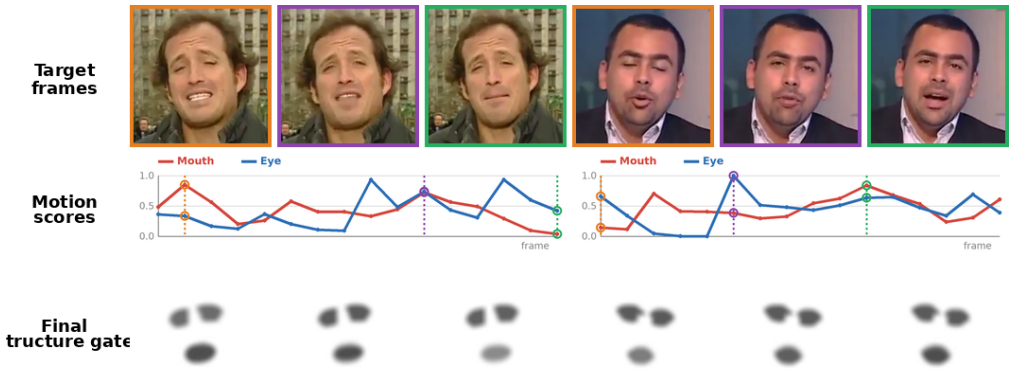


Figure 7: Landmark-guided target-structure preservation. Top: selected target frames from two FF++ clips. Middle: mouth and eye motion scores used to compute the dynamic gate, with colored vertical lines marking the selected frames. Bottom: final structure-gate maps. Darker local regions indicate stronger suppression of the identity residual where target mouth or eye motion should dominate.

Table 8 shows the effect of the shift. At shift 1.0, FVD and expression error are the lowest, but Top-1 identity retrieval collapses to 0.10; the generated videos closely follow the target video's facial motion but do not reliably acquire the source identity. Increasing the shift sharply improves source retrieval, indicating that the identity residual needs sufficient access to higher-noise states to change discriminative facial structure. The final shift-5 setting reaches Top-1/Top-5 retrieval of 0.92/0.98 while retaining comparable expression, pose, and smoothness to the neighboring shift-4 setting. Increasing the shift further to 6.0 gives only a small Top-1 gain and no Top-5 gain, while FVD, expression error, and pose error all increase. We therefore use shift 5.0 as the best identity-quality tradeoff.

Table 9 separates the guidance scale from the sampling schedule. The source guidance s_{src} controls the source-identity editing velocity. When this branch is weak ($s_{\text{src}} = 7.5$), the edit remains conservative and obtains a lower FVD, but source identity transfer is insufficient (Top-1 0.60). Increasing it to 10.5 raises Top-1 retrieval to 0.92, while a further increase to 13.5 brings no identity gain and worsens FVD, indicating over-guidance of the editing branch. The second block varies s_{tar}, which scales the target-identity anchor branch used as the reference velocity in the residual. This setting directly addresses whether the two guidance scales should be matched. With $s_{\text{src}} = 10.5$ fixed, the lower anchor scale $s_{\text{tar}} = 1.5$ is close in identity retrieval but has worse FVD than the default. Matching the scales at $s_{\text{tar}} = 10.5$ is clearly worse: Top-1/Top-5 fall to 0.76/0.94 and expression error rises to 0.5138. Thus, increasing the target-anchor guidance to match the source-identity branch distorts the residual balance between the source-identity velocity and the target-identity anchor. The asymmetric setting $s_{\text{src}} = 10.5$, $s_{\text{tar}} = 3.5$ gives the best overall tradeoff.

Dynamic Protection via Structure Gate. Beyond static semantic modulation, the landmark-guided structure gate explicitly preserves transient facial dynamics over time. As shown in Figure 8, the lower bound ρ_r in Eq. 12 controls the minimum fraction of identity residual retained inside dynamic structural regions. A smaller ρ_r applies stronger suppression, which better preserves mouth and

Shift	Top-1↑	Top-5↑	FVD↓	Smooth↑	Expr↓	Pose↓
1.0	0.10	0.86	127.11	0.9908	0.4141	4.30
2.0	0.60	0.96	220.17	0.9905	0.4741	4.32
2.5	0.70	0.96	248.28	0.9905	0.4842	4.35
4.0	0.90	0.96	307.04	0.9902	0.4908	4.52
5.0	0.92	0.98	313.02	0.9901	0.4939	4.53
6.0	0.94	0.98	321.55	0.9902	0.4964	4.58

Table 8: Sampling-shift ablation on FF++.

s_{src}	s_{tar}	Top-1↑	Top-5↑	FVD↓	Smooth↑	Expr↓	Pose↓
Source guidance variation							
7.5	3.5	0.60	0.96	208.03	0.9910	0.4758	4.38
10.5	3.5	0.92	0.98	313.02	0.9901	0.4939	4.53
13.5	3.5	0.88	0.96	348.59	0.9896	0.4939	4.47
Target guidance variation							
10.5	1.5	0.88	0.96	334.25	0.9901	0.4947	4.58
10.5	3.5	0.92	0.98	313.02	0.9901	0.4939	4.53
10.5	10.5	0.76	0.94	335.62	0.9902	0.5138	4.50

Table 9: Guidance-scale ablation on FF++. The first block varies the source guidance scale while keeping the target guidance scale fixed; the second block varies the target guidance scale while keeping the source guidance scale fixed.

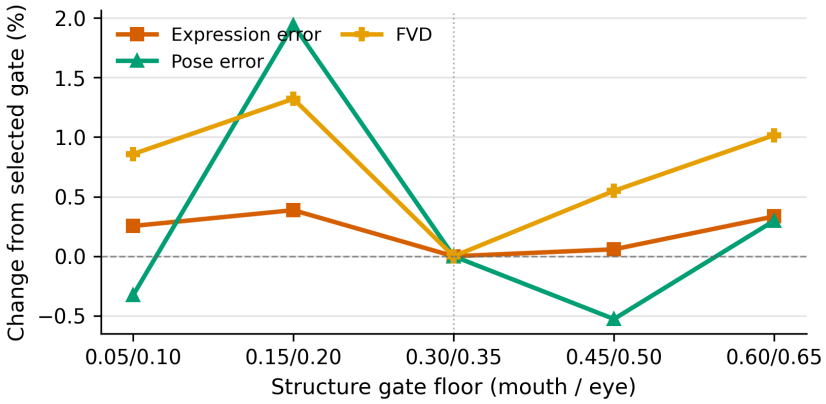


Figure 8: Effect of the mouth/eye structure-gate floor on dynamic preservation and video quality.

eye motion, while a larger ρ_r allows more residual leakage into fast-moving regions, degrading articulation and local visual quality. We use region-specific bounds of $\rho_{\text{mouth}} = 0.30$ and $\rho_{\text{eye}} = 0.35$, which effectively preserve challenging target dynamics.

A.7 Spatial and Temporal Residual Control

The semantic-region visualization in Figure 6 clarifies the spatial role of the residual mask. The identity residual is not uniformly applied over the whole face; it is kept stronger in stable identity regions and attenuated in expression-sensitive or contour-sensitive regions such as the mouth, eyes, and face boundary. This spatial control concentrates identity editing on stable facial areas while reducing residual leakage into regions that are more likely to distort the target expression or silhouette.

Figure 7 illustrates the temporal behavior of our landmark-guided preservation. Unlike the semantic mask, this gate is frame-adaptive: motion scores derived from the target landmark sequence modulate residual suppression in the mouth and eye regions. Stronger target motion yields darker gate values, indicating stronger suppression in these structure-sensitive areas. This ensures that dynamic regions faithfully follow the target's frame-by-frame motion.

A.8 Additional Video Face Swapping Results

Additional video face swapping results are provided in Figure 9. Across both FF++ and CelebVHQ clips, FlowFaceSwap maintains a stable source identity over multiple target frames while following target head pose, mouth articulation, and eye motion. We also include more video examples with the source reference, target video, and corresponding swapped result in the Media Supplement.


Target
Source
Swapped
Target
Source
Swapped
Target
Source
Swapped



Figure 9: Additional video face swapping results.