IDENTITY REFINEMENT WITHOUT RE-EDITING: TRAINING-FREE ENHANCEMENT OF IDENTITY FIDELITY

Anonymous authors
Paper under double-blind review

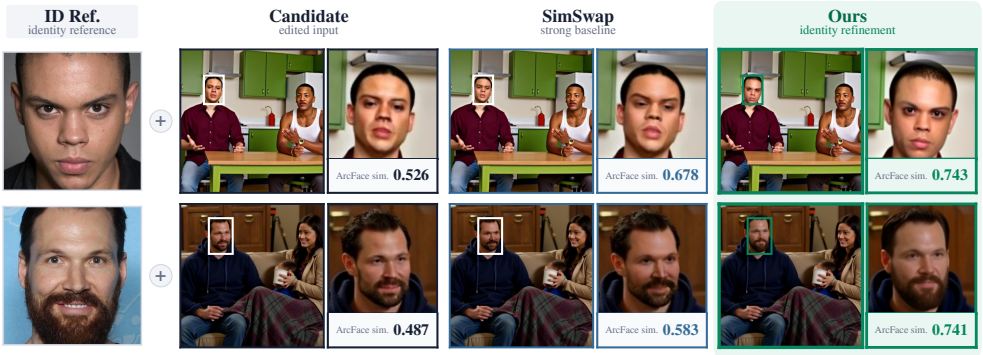


Figure 1: **Post-hoc identity refinement without re-editing.** Starting from an already edited image, our method improves identity fidelity to the reference while preserving completed non-target edit.

ABSTRACT

Person-centric image editing aims to edit an image while preserving the target identity. However, existing identity-preserving methods mainly inject identity information during editing, where identity competes with other editing objectives and can leave residual identity mismatches. Correcting residual identity errors is typically handled by re-editing the whole image, which provides no explicit mechanism for correcting the identity mismatch and may further disturb already satisfactory content. This paper proposes a training-free post-hoc identity refinement framework that improves identity fidelity directly from an edited candidate without re-editing the image from scratch. At each refinement step, the flow state is decomposed into a clean-image endpoint and a complementary noisy-side endpoint. Identity and preservation objectives are applied to the decoded clean endpoint, while the complementary endpoint is retained during exact flow recomposition. To avoid mis-scaling identity guidance for small faces, we introduce support-aware gradient calibration, which estimates gradient magnitude over the facial support. Moreover, a Core–Halo–Background preservation strategy uses region-dependent candidate-trajectory anchoring to protect facial boundaries while retaining already satisfactory non-facial content. The proposed framework is training-free and can be applied across different flow-based models, while achieving SOTA identity refinement and outperforming both training-free and trained identity-preserving baselines. On the MultiID-Bench dataset, the proposed method improves ArcFace, FaceNet, and AdaFace similarity by 16.0%, 12.1%, and 8.1%, respectively, while retaining 99.97% non-target identity and 98.0% outside-target SSIM, achieving the best performance.

1 INTRODUCTION

Modern diffusion- and flow-based image editors can already produce visually compelling edits with substantial control over person placement, clothing, composition, and scene content Kulikov et al. (2025). In parallel, identity-preserving generation has advanced through reference-conditioned

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.

methods such as PhotoMaker Li et al. (2024), InstantID Wang et al. (2024), and PuLID Guo et al. (2024). Yet identity remains more fragile than coarse editing semantics Cheng et al. (2026).This creates a conflict: an edited candidate may already have the desired person placement, body configuration, clothing, background, and surrounding people, while the target face remains insufficiently faithful to the reference. In this setting, regenerating or globally re-editing the image is unnecessary and can reopen degrees of freedom that the upstream editor has already solved. We therefore study *post-hoc identity refinement*: given a fixed edited candidate image and an identity reference image, correct the residual target-identity error while preserving the successful non-target identity content of the candidate image, as shown in Fig. 1.

Existing methods typically address identity while generating or replacing identity-bearing content. Face-swapping methods directly transfer a source identity to a target face Li et al. (2020); Zhao et al. (2023); subject- or reference-conditioned generation methods incorporate identity through subject-specific adaptation or reusable conditioning modules Ruiz et al. (2023); Wang et al. (2024); Guo et al. (2024); and multi-person extensions introduce explicit identity routing or spatial association to reduce identity confusion across people Jiang et al. (2024); Zhang et al. (2025); He et al. (2025). These approaches improve identity during generation, editing, or replacement, placing identity conditioning within the upstream generation process. Our setting begins after this stage: the upstream editor has already produced a semantically acceptable candidate, and the remaining task is to correct residual target-identity error while explicitly preserving what the candidate has already resolved.

Training-free guidance has demonstrated that frozen generative models can be steered at inference time using differentiable external objectives, including recognition-based signals Bansal et al. (2023); Yu et al. (2023); Wallace et al. (2023); He et al. (2024); Ye et al. (2024). Spatial localization and guidance rescaling further provide useful mechanisms for controlling where and how strongly such guidance is applied Avrahami et al. (2022); Mo et al. (2024); Titov et al. (2024). However, directly applying them to post-hoc identity refinement raises **three challenges**:

1. **Trajectory Disruption**: Standard guidance directly modifies the sampling update, which easily distorts the overall image. In contrast, post-hoc refinement needs to keep the existing trajectory intact and only alter the face area.
2. **Scale Sensitivity**: Face identity loss only acts on the face, while background preservation covers the whole image. This mismatch makes the guidance strength heavily depend on face size, leading to unstable results across different scales.
3. **Architectural Coupling**: Existing identity-preserving methods are tied to a specific model architecture. Adapting them to new generative editors requires tedious re-engineering.

To address **trajectory disruption**, we propose a training-free identity refiner that treats the edited candidate as both the initialization basis and the preservation target. *candidate-aligned warm initialization* allocates the larger perturbation budget to the target-face core while keeping non-target regions close to the existing edit. The core mechanism is *flow-consistent clean-endpoint correction*, which decomposes the current flow state into a clean-image endpoint and a complementary (noisy-side) endpoint, applies identity refinement to the decoded clean endpoint, and exactly recomposes the state while retaining the complementary endpoint. This isolates identity correction at an image-interpretable endpoint without backpropagating through the flow model itself.

To address **scale sensitivity**, we introduce *Support-Aware Gradient Calibration (SAGC)* and a *Core–Halo–Background spatial policy*. SAGC calibrates identity guidance according to the effective facial support, avoiding scale-dependent over- or under-guidance. The Core–Halo–Background policy separates editable regions, boundary protection, and reliable background anchoring. Most importantly, *the proposed refinement logic is decoupled from the backbone topology*, addressing the challenge of **architectural coupling**. These refinement components are external to the generative backbone, requiring only backbone-specific flow prediction and state recomposition, which enables application across different flow-based editors, including FLUX-style and Qwen-style architectures.

Extensive experiments across multiple identity recognizers and flow-based editors demonstrate consistent improvements in identity fidelity while preserving the completed edit. On MultiID-Bench Xu et al. (2026), it is the only approach with positive mean identity gain across all three recognizers, while additional ablation studies validate the proposed guidance, spatial preservation, and cross-backbone generalization. Our main contributions are summarized as follows:

- We formulate identity correction as a **post-hoc refinement problem over an already edited candidate** and introduce flow-consistent clean-endpoint correction, which updates the decoded clean endpoint while retaining the complementary endpoint and exactly recomposing the linear-flow state. This improves residual identity fidelity while keeping already-resolved candidate content stable.
- We introduce **Support-Aware Gradient Calibration (SAGC)** and a **Core–Halo–Background policy** to address localized guidance scaling and spatial preservation. SAGC calibrates identity gradients over their effective facial support, while the halo protects the editable boundary without weakening anchoring in the farther background.
- We **decouple identity refinement from the underlying editor backbone**, restricting architecture-specific operations to endpoint recovery and recomposition while keeping identity guidance, preservation, calibration, and spatial control external. This yields a **backbone-agnostic post-hoc refinement framework** that operates across structurally different compatible flow-based editors.

2 Related Work

2.1 Identity-Preserving Person Generation and Editing

Face swapping transfers a source identity to a target face while retaining selected target attributes: FaceShifter learns identity-aware replacement with occlusion handling Li et al. (2020), SimSwap uses generalized identity injection Chen et al. (2020), and DiffSwap introduces 3D-aware masked diffusion Zhao et al. (2023). Subject-driven personalization incorporates reference appearance through model or conditioning adaptation, including DreamBooth Ruiz et al. (2023) and Custom Diffusion Kumari et al. (2023); tuning-free approaches such as FastComposer Xiao et al. (2025), IP-Adapter Ye et al. (2023), PhotoMaker Li et al. (2024), Face2Diffusion Shiohara & Yamasaki (2024), Infinite-ID Wu et al. (2024), InstantID Wang et al. (2024), and PuLID Guo et al. (2024) provide reusable identity conditioning without per-subject optimization.

Multi-person methods extend identity conditioning with explicit routing and spatial association. GroupDiff Jiang et al. (2024), MS-Diffusion Wang et al. (2025b), MultiBooth Zhu et al. (2025), ID-Patch Zhang et al. (2025), UniPortrait He et al. (2025), DynamicID Hu et al. (2025), and PersonaCraft Kim et al. (2025) combine identity features with region, layout, pose, or routing controls, while WithAnyone Xu et al. (2026), GroupPortrait Huang et al. (2026), PositionIC Hu et al. (2026), and UMO Cheng et al. (2026) further address multi-person consistency, identity–position assignment, and cross-person identity confusion. Across these families, identity is incorporated while identity-bearing content is generated, edited, or replaced. Identity refinement, however, targets the subsequent stage: an upstream editor has already produced a semantically acceptable candidate, and refinement focuses on the residual target-identity error under explicit preservation of the candidate content that is already correct. Blind face restoration addresses a complementary setting, with GFP-GAN Wang et al. (2021) and CodeFormer Zhou et al. (2022) recovering plausible facial detail from degraded observations rather than conditioning reconstruction on a separate identity reference.

2.2 Training-Free Guidance, Spatial Control, and Flow-Based Editing

Training-free guidance enables pretrained generative models to optimize external objectives during inference. Universal Guidance Bansal et al. (2023) and FreeDoM Yu et al. (2023) steer frozen diffusion models with differentiable objectives, including recognition-based signals; DOODL Wallace et al. (2023) and MPGD He et al. (2024) further integrate differentiable guidance into the generative trajectory, and TFG Ye et al. (2024) provides a unified view of training-free guidance and highlights the importance of how external gradients are injected into the sampling process. Spatially controlled methods such as Blended Diffusion Avrahami et al. (2022), DiffEdit Couairon et al. (2023), MAG-Edit Mao et al. (2024), LIME Simsar et al. (2025), and FreeControl Mo et al. (2024) localize editing through masks, attention constraints, or spatial controls. Guide-and-Rescale Titov et al. (2024) and GradNorm Chen et al. (2018) further motivate explicit calibration of heterogeneous guidance signals. Together, these works establish external differentiable guidance, spatial localization, and gradient rescaling as effective ingredients for inference-time control. They do not address the

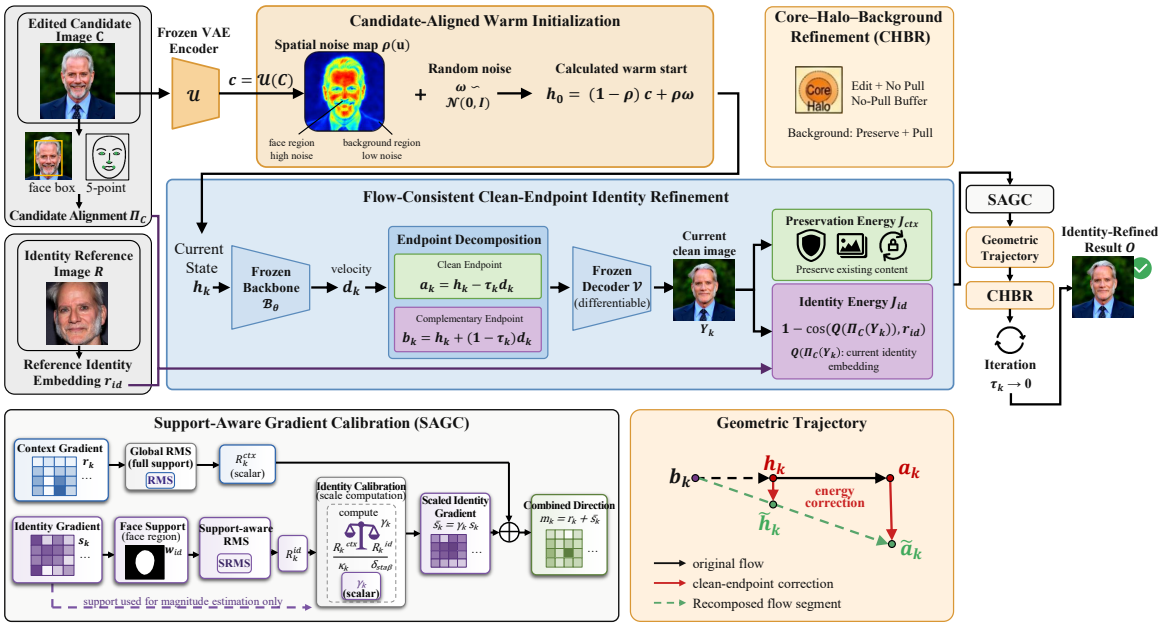


Figure 2: **Overview of the proposed training-free post-hoc identity refinement framework.** Given an edited candidate and an identity reference, we apply candidate-aligned warm initialization to allocate perturbation budget strictly to the target face. At each refinement step, the latent flow state is decomposed into clean and complementary endpoints; external identity and preservation gradients update only the decoded clean endpoint, followed by exact flow recomposition. Support-Aware Gradient Calibration (SAGC) stabilizes face-localized guidance, while a Core–Halo–Background policy anchors identity-irrelevant context.

support-size dependence that arises when a highly localized facial identity gradient is normalized over a much larger latent field.

Beyond generic inference-time guidance, flow-based editing introduces a separate question of how external corrections should be transported through the generative trajectory. Flow matching has motivated editing and inversion mechanisms that respect the transport structure of rectified-flow models Lipman et al. (2023). RF-Inversion Rout et al. (2025), RF-Solver/RF-Edit Wang et al. (2025a), FlowEdit Kulikov et al. (2025), and FlowAlign Kim et al. (2026) recover or manipulate source content through flow-specific trajectories and update rules. RewardFlow Susladkar et al. (2026) and training-free trajectory optimal control Chang et al. (2026) further combine differentiable rewards with diffusion or flow models. However, Residual identity refinement poses conflict: identity reward can interfere with the native flow transport and background preservation can suppress the required identity correction. This conflict motivate transport-compatible and spatially selective refinement.

3 Methodology

Let $\mathbf{C}$ be an edited candidate that already satisfies the intended semantics and $\mathbf{R}$ an identity reference. We seek $\mathbf{O} = \mathcal{R}(\mathbf{C}, \mathbf{R})$ whose target identity better matches $\mathbf{R}$ while retaining the layout, interactions, and non-target content of $\mathbf{C}$. The upstream editor is outside our formulation; the candidate is both the initialization basis and the preservation target. Fig. 2 summarizes the pipeline.

3.1 Candidate-aligned initialization.

With frozen latent encoder $\mathcal{U}$, let $\mathbf{c} = \mathcal{U}(\mathbf{C})$ be the candidate latent, $\mathbf{u}$ a spatial location, and $\chi_{\text{core}}(\mathbf{u}) \in \{0, 1\}$ the target-face-core indicator. We define the spatial perturbation map $\rho(\mathbf{u})$ and warm-start latent $\mathbf{h}_0(\mathbf{u})$ as

$$\rho(\mathbf{u}) = \rho_{\text{keep}} + \chi_{\text{core}}(\mathbf{u})(\rho_{\text{edit}} - \rho_{\text{keep}}), \tag{1}$$

$$\mathbf{h}_0(\mathbf{u}) = [1 - \rho(\mathbf{u})]\mathbf{c}(\mathbf{u}) + \rho(\mathbf{u})\boldsymbol{\omega}(\mathbf{u}), \quad \boldsymbol{\omega} \sim \mathcal{N}(0, I), \tag{2}$$

where $\rho_{\text{edit}} > \rho_{\text{keep}}$ are the perturbation strengths inside and outside the face core, respectively. Thus, the face core receives a larger search radius while already-correct regions stay near the candidate. The target face and its five-point similarity transform $\Pi_{\mathbf{C}}$ are determined once from $\mathbf{C}$ and kept fixed throughout refinement to maintain stable alignment and avoid step-wise landmark drift.

3.2 Flow-Consistent Clean-Endpoint Correction

At refinement step k, let $\mathbf{h}_k$ denote the current latent state, τ_k the decreasing flow coordinate, and $\boldsymbol{\psi}$ the fixed backbone conditioning. The frozen flow backbone $\mathcal{B}_\vartheta$ predicts the displacement

$$\mathbf{d}_k = \mathcal{B}_\vartheta(\mathbf{h}_k, \tau_k; \boldsymbol{\psi}), \tag{3}$$

where ϑ denotes its frozen parameters. Under the linear-flow parameterization

$$\mathbf{h}_k = (1-\tau_k)\mathbf{a}_k + \tau_k \mathbf{b}_k, \qquad \mathbf{d}_k = \mathbf{b}_k - \mathbf{a}_k, \tag{4}$$

the clean endpoint $\mathbf{a}_k$ and complementary (noisy-side) endpoint $\mathbf{b}_k$ are recovered exactly:

$$\mathbf{a}_k = \mathbf{h}_k - \tau_k \mathbf{d}_k, \tag{5}$$

$$\mathbf{b}_k = \mathbf{h}_k + (1-\tau_k)\mathbf{d}_k. \tag{6}$$

Our key operation modifies only $\mathbf{a}_k$ while keeping $\mathbf{b}_k$ fixed.

Decoded objectives. Let $\mathbf{Y}_k = \mathcal{V}(\mathbf{a}_k)$ be the decoded clean estimate. We use a candidate-retention energy

$$\mathcal{J}_{\text{ctx}}(\mathbf{a}_k) = \frac{1}{2}\sum_{\mathbf{u}} \varpi(\mathbf{u}) \|\mathbf{Y}_k(\mathbf{u}) - \mathbf{C}(\mathbf{u})\|_2^2, \tag{7}$$

where ϖ is small in the editable core and large in reliable regions, and an identity energy

$$\mathcal{J}_{\text{id}}(\mathbf{a}_k) = 1 - \cos[\mathcal{Q}(\Pi_{\mathbf{C}}(\mathbf{Y}_k)), \mathbf{r}_{\text{id}}], \quad \mathbf{r}_{\text{id}} = \mathcal{Q}(\Pi_{\mathbf{R}}(\mathbf{R})). \tag{8}$$

Both are differentiated through the frozen decoder. After calibration (Sec. 3.3), let $\mathcal{H}_k = u_k \mathcal{J}_{\text{ctx}} + s_k \mathcal{J}_{\text{id}}$, a short backtracking search proposes $\mathbf{a}_k^\dagger = \mathbf{a}_k - \beta \mathbf{m}_k$ and accepts the first trial satisfying

$$\mathcal{H}_k(\mathbf{a}_k^\dagger) \le \mathcal{H}_k(\mathbf{a}_k) + \delta_{\text{num}}, \tag{9}$$

$$\mathcal{J}_{\text{id}}(\mathbf{a}_k^\dagger) \le \mathcal{J}_{\text{id}}(\mathbf{a}_k) + \delta_{\text{num}}; \tag{10}$$

otherwise the original endpoint is retained. Full pseudocode is in Appendix F (Alg. 1).

Exact flow recomposition. The accepted endpoint $\bar{\mathbf{a}}_k$ is recomposed with the unchanged complementary endpoint:

$$\widetilde{\mathbf{h}}_{k+1} = (1-\tau_{k+1})\bar{\mathbf{a}}_k + \tau_{k+1}\mathbf{b}_k. \tag{11}$$

Let $\Delta\mathbf{a}_k = \bar{\mathbf{a}}_k - \mathbf{a}_k$ and $\mathbf{E}_k = \mathbf{h}_k + (\tau_{k+1} - \tau_k)\mathbf{d}_k$ denote the accepted clean correction and native Euler state. Equation 11 gives $\widetilde{\mathbf{h}}_{k+1} = \mathbf{E}_k + (1-\tau_{k+1})\Delta\mathbf{a}_k$. This exact endpoint transport defines the flow consistency used here; at $\Delta\mathbf{a}_k = \mathbf{0}$, recomposition recovers the native deterministic Euler step. The regional preservation pull in Eq. 15 is applied afterward. Appendix G.2 derives matched transport alternatives and the corresponding learned-field diagnostic.

3.3 Stable Localized Refinement

Support-Aware Gradient Calibration. Let $\mathbf{r}_k = \nabla_{\mathbf{a}_k}[u_k \mathcal{J}_{\text{ctx}}]$ and $\mathbf{s}_k = \nabla_{\mathbf{a}_k}\mathcal{J}_{\text{id}}$. For soft facial support $\mathbf{w}_{\text{id}}$, define

$$\text{SRMS}(\mathbf{z}; \mathbf{w}) = \sqrt{\frac{\sum_{i,c} w_i z_{i,c}^2}{C\sum_i w_i}}, \qquad \gamma_k = \frac{\kappa_k \,\text{RMS}(\mathbf{r}_k)}{\text{SRMS}(\mathbf{s}_k; \mathbf{w}_{\text{id}}) + \delta_{\text{stab}}}, \qquad \mathbf{m}_k = \mathbf{r}_k + \gamma_k \mathbf{s}_k. \tag{12}$$

The support is used only to estimate magnitude, while the full identity gradient is propagated through the decoder receptive field. For a fixed gradient, let $p = N^{-1}\sum_i w_i$ and $q = \sum_{i,c} w_i s_{i,c}^2 / \sum_{i,c} s_{i,c}^2$. Then $\text{SRMS}(\mathbf{s}; \mathbf{w})/\,\text{RMS}(\mathbf{s}) = \sqrt{q/p}$. When the facial support contains most identity-gradient energy, $q \simeq 1$, so whole-latent calibration applies approximately $1/\sqrt{p}$ more gain at a fixed target magnitude; SAGC calibrates directly with the support-conditioned denominator (Appendix G.3).

Core–Halo–Background refinement. We use $\Omega_{\text{core}} \subset \Omega_{\text{halo}}$. The core receives the larger initialization budget; the halo is a no-pull buffer for nearby decoder responses; outside the halo, the candidate is treated as reliable. At the next flow coordinate, define

$$\mathbf{q}_{k+1} = (1 - \bar{\rho}_{k+1})\mathbf{c} + \bar{\rho}_{k+1}\boldsymbol{\omega}, \quad \bar{\rho}_{k+1} = \rho_{\text{keep}} \frac{\tau_{k+1}}{\tau_0}, \tag{13}$$

$$\pi(\mathbf{u}) = \zeta[1 - \chi_{\text{halo}}(\mathbf{u})], \tag{14}$$

$$\mathbf{h}_{k+1}(\mathbf{u}) = [1 - \pi(\mathbf{u})]\widetilde{\mathbf{h}}_{k+1}(\mathbf{u}) + \pi(\mathbf{u})\mathbf{q}_{k+1}(\mathbf{u}). \tag{15}$$

Hence the core has edit freedom, the halo shields the editable boundary from direct trajectory-pull interference, and the farther background is contracted toward the candidate trajectory with the full pull coefficient; the contraction property is given in Appendix H.4. Appendix D.3 isolates this spatial role with matched pre-pull states and a global-intervention-norm-matched control.

3.4 Backbone-Decoupled Instantiation

The refiner only requires a latent state, a backbone prediction from which the two endpoints are recoverable, a differentiable decoder, and a valid recomposition rule. Backbone-specific logic is therefore confined to $\mathcal{B}_\vartheta, \mathcal{U}, \mathcal{V}$, and conditioning ψ; the objectives, SAGC, backtracking, and spatial policy remain unchanged. We instantiate this interface on a FLUX-style joint-attention model (Black Forest Labs, 2026) and a Qwen-style causal-reference model (LightX2V Team, 2026), which substantially differ internally but expose the same clean-endpoint refinement interface.

4 Experiments

4.1 Experimental Setup

Protocol. We evaluate on MultiID-Bench (Xu et al., 2026). The fixed-candidate benchmark contains 707 source–target records; after three result-blind candidate-localization exclusions, 704 remain, split by source image into 64 calibration and 640 untouched test targets. Common candidates are generated once with the FLUX.1-dev-based WithAnyone model and then frozen together with target references and evaluation geometry; failed method outputs fall back to the unchanged candidate and remain in the denominator. For cross-source evaluation, all 227 multi-person tasks (501 references) are attempted by four candidate generators, after which candidate faces are Hungarian-matched to references and the assignment is frozen before refinement. Additional protocol details, denominators, seeds, and hyperparameters are in Appendix A.

Metrics and implementation. Target identity is measured with ArcFace (Deng et al., 2019), FaceNet (Schroff et al., 2015), and AdaFace (Kim et al., 2022); we report gain over the frozen candidate and positive-gain win rate. Preservation is evaluated against the candidate using non-target identity retention and outside-target image similarity (RGB ℓ_1/SSIM (Wang et al., 2004) in the fixed-candidate benchmark; LPIPS (Zhang et al., 2018) elsewhere). All networks are frozen. Experiments use 512×512 images on an 80 GB A100 and four refinement steps with FLUX.2 [klein] 9B-KV (Black Forest Labs, 2026) or Qwen-Image-Edit-Causal (LightX2V Team, 2026).

4.2 Comparison with Other Methods

We compare with native FLUX.2 (Black Forest Labs, 2026), Qwen-Image-Edit-2511 (Qwen Team, 2025), WithAnyone Kontext (Xu et al., 2026), SimSwap (Chen et al., 2020), REFace (Baliah et al., 2025), Face-Adapter (Han et al., 2024), and APPLE (Kang et al., 2026), plus the unchanged candidate. Crop-based methods use the same frozen target localization.

Tab. 1 shows that our method is the only evaluated approach with positive mean gains on all three recognizers and achieves the largest gain for each. Against native FLUX.2, outside-region ℓ_1 drops from 0.057835 to 0.000002 and SSIM rises from 0.81664 to 0.99996, while non-target identity remains 0.99994. Compared with SimSwap, the strongest alternative on FaceNet and AdaFace, our gains are higher by 0.0180 on both metrics; source-clustered 95% confidence intervals are $[0.0082, 0.0276]$ and $[0.0008, 0.0350]$, respectively.

Table 1: Fixed-candidate comparison on MultiID-Bench. NT error and outside-region errors measure deviations from the frozen candidate and are lower-is-better. The best and second identity gains are **bolded** and underlined, excluding the unchanged candidate.

Method	ΔArcFace ↑	ΔFaceNet ↑	ΔAdaFace ↑	NT error ↓	Outside ℓ_1 ↓	Outside SSIM error ↓	Fallbacks
Candidate	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0
FLUX.2	-0.0953	+0.0248	-0.1053	6.4420e-02	5.7835e-02	1.8336e-01	0
Qwen-Image-Edit	-0.1998	-0.0315	-0.2285	1.5000e-04	3.8500e-04	1.6800e-03	0
WithAnyone (Kontext)	-0.0118	+0.0039	-0.0283	2.1000e-04	2.0000e-05	**2.0000e-05**	1
SimSwap	-0.0169	+0.0675	+0.0263	2.2540e-02	4.4000e-04	1.9700e-03	22
REFace	-0.1810	-0.0712	-0.1567	1.1000e-04	1.0740e-03	4.9400e-03	21
Face-Adapter	-0.0685	+0.0307	-0.0482	3.1000e-04	5.8000e-05	1.2000e-04	8
APPLE	-0.0311	+0.0084	-0.0362	7.0000e-05	9.0000e-06	5.0000e-05	22
Ours (FLUX.2)	**+0.0969**	**+0.0856**	**+0.0443**	**6.0000e-05**	**2.0000e-06**	4.00000e-05	0



Figure 3: Qualitative comparison on common edited candidates. Each row shows the identity reference, frozen candidate, seven baselines, and our result; footer values are target ArcFace cosine similarities.

Fig. 3 confirms the quantitative trend across portraits and multi-person scenes: our outputs better match the target reference while retaining clothing, body pose, neighboring people, and scene structure already established by the candidate. A step-wise clean-endpoint diagnostic is provided in Appendix B.

4.3 Ablation Study

We ablate the main design choices on the same 494 FLUX.2 targets with candidates, references, assignments, weights, and seeds fixed. Tab. 2 summarizes the central comparisons; full metrics, support-stratified analysis, and qualitative ablations are in Appendix C.

SAGC reduces support-size-dependent calibration bias: compared with global RMS, it improves FaceNet transfer, outside-region preservation, and non-target retention while providing consistent positive identity improvement across all targets. Candidate-aligned initialization raises AdaFace gain from 0.2447 to 0.3056 while reducing outside LPIPS by 36.1%. Fixed candidate coordinates outperform dynamic alignment, which also incurs one landmark failure. For the spatial policy, core-only editing increases identity gain but causes substantial outside drift; adding candidate-trajectory pull reduces LPIPS from 0.03256 to 0.01196.

Table 2: Compact ablation summary. Higher identity gains and NT are better; lower outside LPIPS is better. Full win rates and diagnostic metrics are reported in Appendix C.

Configuration	ΔAdaFace ↑	ΔFaceNet ↑	Outside LPIPS ↓	NT ↑
Gradient calibration				
Global RMS	+0.3655	+0.1139	0.01472	0.9724
SAGC (Ours)	+0.3056	**+0.1314**	**0.01257**	**0.9738**
Candidate alignment				
Native full-noise restart	+0.2447	+0.0885	0.01967	0.9640
Candidate-aligned (Ours)	**+0.3056**	**+0.1314**	**0.01257**	**0.9738**
Dynamic five-point	+0.2722	+0.1174	**0.01254**	0.9719
Fixed five-point (Ours)	**+0.3058**	**+0.1313**	0.01257	**0.9738**
Spatial policy				
Uniform	+0.0763	+0.0356	**0.00659**	0.9589
Core only	+0.2654	+0.1152	0.03256	0.9453
Core + pull (Ours)	**+0.3064**	**+0.1314**	0.01196	**0.9757**

Table 3: Matched-state effect of halo protection. Lower is better.

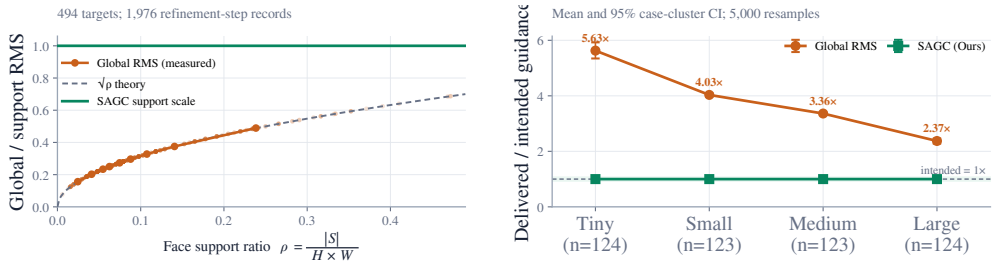
Metric	No halo	Halo	Reduction
Boundary pull RMS	0.010645	**0.004810**	**54.8%**
BPGJ	0.009249	**0.005301**	**42.7%**
Boundary LPIPS	0.001963	**0.000809**	**58.8%**

The halo component is evaluated separately through a matched-state intervention. As shown in Table 3, a matched-state intervention on 195 candidate images reduces boundary pull RMS by 54.8%, boundary gradient interference by 42.7%, and boundary LPIPS by 58.8%. Further norm-matched controls are provided in Appendix D.3.

Controlled guidance and transport. On a source-image split of the same 494-target FLUX.2 evaluation, task-matched guidance controls share the candidate/reference inputs, ArcFace and preservation objectives, SAGC, Core–Halo–Background policy, four-step schedule, and seed. On 441 held-out targets, our canonical update raises FaceNet gain from 0.0989 to 0.1319 relative to mixed TFG guidance and lowers outside LPIPS from 0.01664 to 0.01265; paired 95% intervals for the differences are $[0.0212, 0.0457]$ and $[-0.00575, -0.00242]$. Relative to state guidance, AdaFace/FaceNet gains increase from $0.0885/0.0370$ to $0.3078/0.1319$. The endpoint update uses four backbone predictions and zero backbone backward passes, with 38.75 GiB peak allocation and 2.68 s synchronized core latency versus approximately 77.2 GiB and 4.38–4.41 s for the two trajectory-guidance controls (see Appendix D.1 Tab. 12).

With the optimization pipeline fixed and each transport independently calibrated, fixed-complementary-endpoint recomposition improves AdaFace gain, FaceNet gain, and outside LPIPS over fixed-current-state clean/velocity conversion ($0.3753/0.1504/0.01147$ versus $0.3005/0.1162/0.01312$). Endpoint recomposition improves FaceNet transfer over alternative transport rules while maintaining competitive identity gains and preservation. The acceptance guard further supplies a clean-step safeguard: at strength 4, guarded endpoint correction retains FaceNet gain $+0.0930$ with LPIPS 0.01288, whereas the unguarded arm reaches -0.0575 and 0.08541; over 1,764 held-out steps at the selected strength, the unguarded arm admits 275 identity-loss-increasing clean trials and the guarded arm admits zero. Full transport curves, paired intervals, acceptance counts, and local learned-field diagnostics are in Appendix D.2.

Fig. 4 analyzes 1,976 step records from the 494 trajectories. The global-to-support normalization ratio grows as effective support shrinks, consistent with the $\sqrt{q/p}$ support-energy relation in Sec. 3.3: counterfactual global calibration delivers $5.63\times$ the intended local guidance for the smallest face group versus $2.37\times$ for the largest. On the 494 final outputs, SAGC increases FaceNet gain over global RMS by $+0.01746$ (paired 95% CI $[+0.00769, +0.02721]$), while improving outside LPIPS by 0.00215 and non-target retention by 0.00133; global RMS retains larger AdaFace and guidance-ArcFace gains. Full quartile and calibration diagnostics are in Appendix C.



(a) Support-size bias of global RMS **(b) Delivered identity guidance**

Figure 4: Support-aware gradient calibration. The relation between support RMS and whole-latent RMS depends on effective support fraction and in-support gradient energy.

Table 4: Cross-source/backbone evaluation on matched targets. Win is the AdaFace positive-gain rate; NT is non-target AdaFace retention.

Candidate source	Refiner	ΔAdaFace ↑	Win (%) ↑	Outside LPIPS ↓	NT ↑	N
FLUX.2	FLUX.2	+0.3056	100.00	0.01257	0.9738	494
	Qwen	+0.1335	83.20	0.02995	0.8925	494
Qwen-Image-Edit	FLUX.2	+0.2695	100.00	0.01294	0.9800	477
	Qwen	+0.1479	94.97	0.01290	0.9575	477
OmniGen2	FLUX.2	+0.3521	99.80	0.01334	0.9748	489
	Qwen	+0.1555	88.55	0.02622	0.9332	489
WithAnyone	FLUX.2	+0.1623	99.79	0.01361	0.9835	483
	Qwen	+0.0450	74.95	0.02433	0.9531	483

4.4 Generalization across Sources and Backbones

In Tab. 4, we independently vary the upstream candidate generator (FLUX.2, Qwen-Image-Edit-2511, OmniGen2 (Wu et al., 2025), WithAnyone) and the refinement backbone (FLUX.2 or Qwen). Within each source, both refiners receive identical candidate pixels, references, and assignments.

All eight pairings have positive mean identity gains under all three recognizers (full matrix in Appendix E). FLUX.2 yields AdaFace gains of 0.1623–0.3521 with 99.79–100% win rates; Qwen yields 0.0450–0.1555 with 74.95–94.97%. Even the smallest mean gain has a case-clustered 95% interval above zero, $[0.0370, 0.0532]$. Thus, both structurally different backbones support the same post-hoc refinement formulation. Additional source coverage, qualitative results, and runtime/memory measurements are in Appendix E.

5 Conclusion

In this paper, we formulate residual identity correction after person-centric image editing as a training-free post-hoc refinement problem, where an already satisfactory candidate is retained while its remaining target-identity error is corrected. This post-hoc formulation is complementary to existing identity-preserving techniques and can be composed with them as an additional refinement stage. Our framework combines flow-consistent clean-endpoint correction with exact linear-flow recomposition, support-aware gradient calibration, and candidate-relative spatial protection, while keeping the refinement logic decoupled from the backbone. On MultiID-Bench, our method is the only one among nine evaluated approaches with positive mean identity gain across ArcFace, FaceNet, and AdaFace. Held-out guidance comparisons and matched-state interventions further validate the proposed correction and spatial-preservation mechanisms, with halo protection reducing boundary pull interference by 54.8%. Across four candidate sources and two structurally different refinement backbones, all eight source–refiner combinations achieve positive mean identity metric gains. In summary, these results establish post-hoc refinement as an effective way to improve residual identity fidelity while preserving the non-target content of an existing edit. Future work will investigate generation-oriented identity objectives that better align identity fidelity during image generation.

AI USE STATEMENT

General-purpose generative AI assistants were used to assist with code writing and language polishing. The authors reviewed all AI-assisted content and take full responsibility for the paper.

REFERENCES

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18187–18197, 2022. doi: 10.1109/CVPR52688.2022.01767. URL https://doi.org/10.1109/CVPR52688.2022.01767.

Sanoojan Baliah, Qinliang Lin, Shengcai Liao, Xiaodan Liang, and Muhammad Haris Khan. Realistic and efficient face swapping: A unified approach with diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Baliah_Realistic_and_Efficient_Face_Swapping_A_Unified_Approach_with_Diffusion_WACV_2025_paper.html.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 843–852, 2023. doi: 10.1109/CVPRW59228.2023.00091. URL https://doi.org/10.1109/CVPRW59228.2023.00091.

Black Forest Labs. FLUX.2 [klein] 9B-KV. Official model release, 2026. URL https://huggingface.co/black-forest-labs/FLUX.2-klein-9b-kv. Accessed September 5, 2026.

Jinho Chang, Jaemin Kim, and Jong Chul Ye. Training-free reward-guided image editing via trajectory optimal control, 2026. URL https://doi.org/10.48550/arXiv.2509.25845. International Conference on Learning Representations (ICLR).

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*, pp. 2003–2011, 2020. doi: 10.1145/3394171.3413630. URL https://doi.org/10.1145/3394171.3413630.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks, 2018. URL https://doi.org/10.48550/arXiv.1711.02257.

Yufeng Cheng, Wenxu Wu, Shaojin Wu, Mengqi Huang, Fei Ding, and Qian He. Umo: Scaling multi-identity consistency for image customization via matching reward, 2026. URL https://doi.org/10.48550/arXiv.2509.06818. CVPR 2026; canonical preprint DOI used pending an indexed proceedings DOI.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *International Conference on Learning Representations*, 2023.

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*, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html.

Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1159. URL https://doi.org/10.52202/079017-1159.

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*, 2024. URL https://arxiv.org/abs/2405.12970.

Junjie He, Yifeng Geng, and Liefeng Bo. Uniportrait: A unified framework for identity-preserving single- and multi-human image personalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14399–14408, 2025. doi: 10.1109/ICCV51701.2025.01336. URL https://doi.org/10.1109/ICCV51701.2025.01336.

Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J. Zico Kolter, Ruslan Salakhutdinov, and Stefano Ermon. Manifold preserving guided diffusion, 2024. URL https://doi.org/10.48550/arXiv.2311.16424. International Conference on Learning Representations (ICLR).

Junjie Hu, Tianyang Han, Kai Ma, Jialin Gao, Hao Dou, Song Yang, Xianhua He, Jianhui Zhang, Junfeng Luo, Xiaoming Wei, and Wenqiang Zhang. Positionic: Unified position and identity consistency for image customization, 2026. URL https://doi.org/10.48550/arXiv.2507.13861. CVPR 2026; canonical preprint DOI used pending an indexed proceedings DOI.

Xirui Hu, Jiahao Wang, Hao Chen, Weizhan Zhang, Benqi Wang, Yikun Li, and Haishun Nan. Dynamicid: Zero-shot multi-id image personalization with flexible facial editability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10549–10559, 2025. doi: 10.1109/ICCV51701.2025.00982. URL https://doi.org/10.1109/ICCV51701.2025.00982.

Meijia Huang, Ruida Li, Bing Ma, Liangwei Jiang, Shuo Fang, and Chenguang Ma. Grouppportrait: Multi-id portrait generation with high identity preservation and fine-grained control. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6226–6235, 2026. doi: 10.1109/WACV61042.2026.00602. URL https://doi.org/10.1109/WACV61042.2026.00602.

Yuming Jiang, Nanxuan Zhao, Qing Liu, Krishna Kumar Singh, Shuai Yang, Chen Change Loy, and Ziwei Liu. Groupdiff: Diffusion-based group portrait editing. In *Computer Vision – ECCV 2024*, pp. 221–239, 2024. doi: 10.1007/978-3-031-72754-2_13. URL https://doi.org/10.1007/978-3-031-72754-2_13.

Jiwon Kang, Yeji Choi, JoungBin Lee, Wooseok Jang, Jinhyeok Choi, Taekeun Kang, Yongjae Park, Myungin Kim, and Seungryong Kim. APPLE: Attribute-preserving pseudo-labeling for diffusion-based face swapping, 2026. URL https://arxiv.org/abs/2601.15288. CVPR 2026.

Gwanghyun Kim, Suh Yoon Jeon, Seunggyu Lee, and Se Young Chun. Personacraft: Personalized and controllable full-body multi-human scene generation using occlusion-aware 3d-conditioned diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12034–12044, 2025. doi: 10.1109/ICCV51701.2025.01119. URL https://doi.org/10.1109/ICCV51701.2025.01119.

Jeongsol Kim, Yeobin Hong, Jonghyun Park, and Jong Chul Ye. Flowalign: Trajectory-regularized, inversion-free flow-based image editing, 2026. URL https://doi.org/10.48550/arXiv.2505.23145. International Conference on Learning Representations (ICLR).

Minchul Kim, Anil K. Jain, and Xiaoming Liu. AdaFace: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Kim_AdaFace_Quality_Adaptive_Margin_for_Face_Recognition_CVPR_2022_paper.html.

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*, pp. 19721–19730, 2025. doi: 10.1109/ICCV51701.2025.01834. URL https://doi.org/10.1109/ICCV51701.2025.01834.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023. doi: 10.1109/CVPR52729.2023.00192. URL https://doi.org/10.1109/CVPR52729.2023.00192.

Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5074–5083, 2020. doi: 10.1109/CVPR42600.2020.00512. URL https://doi.org/10.1109/CVPR42600.2020.00512.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8640–8650, 2024. doi: 10.1109/CVPR52733.2024.00825. URL https://doi.org/10.1109/CVPR52733.2024.00825.

LightX2V Team. Qwen-Image-Edit-Causal. Official model release, 2026. URL https://huggingface.co/lightx2v/Qwen-Image-Edit-Causal. Accessed September 5, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://doi.org/10.48550/arXiv.2210.02747.

Qi Mao, Lan Chen, Yuchao Gu, Zhen Fang, and Mike Zheng Shou. Mag-edit: Localized image editing in complex scenarios via mask-based attention-adjusted guidance. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6842–6850, 2024. doi: 10.1145/3664647.3680830.

Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7465–7475, 2024. doi: 10.1109/CVPR52733.2024.00713. URL https://doi.org/10.1109/CVPR52733.2024.00713.

Qwen Team. Qwen-Image-Edit-2511. Official model release, 2025. URL https://huggingface.co/Qwen/Qwen-Image-Edit-2511. Accessed September 5, 2026.

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using stochastic rectified differential equations, 2025. URL https://doi.org/10.48550/arXiv.2410.10792. International Conference on Learning Representations (ICLR).

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023. doi: 10.1109/CVPR52729.2023.02155. URL https://doi.org/10.1109/CVPR52729.2023.02155.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. doi: 10.1109/CVPR.2015.7298682. URL https://arxiv.org/abs/1503.03832.

Kaede Shiohara and Toshihiko Yamasaki. Face2diffusion for fast and editable face personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6850–6859, 2024. doi: 10.1109/CVPR52733.2024.00654. URL https://doi.org/10.1109/CVPR52733.2024.00654.

Enis Simsar, Alessio Tonioni, Yongqin Xian, Thomas Hofmann, and Federico Tombari. Lime: Localized image editing via attention regularization in diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 222–231, 2025. doi: 10.1109/WACV61041.2025.00032.

Onkar Susladkar, Dong-Hwan Jang, Tushar Prakash, Adheesh Juvekar, Vedant Shah, Ayush Barik, Nabeel Bashir, Muntasir Wahed, Ritish Shrirao, and Ismini Lourentzou. Rewardflow: Generate images by optimizing what you reward, 2026. URL https://doi.org/10.48550/arXiv.2604.08536. CVPR 2026; canonical preprint DOI used pending an indexed proceedings DOI.

Vadim Titov, Madina Khalmatova, Alexandra Ivanova, Dmitry Vetrov, and Aibek Alanov. Guide-and-rescale: Self-guidance mechanism for effective tuning-free real image editing. In *Computer Vision – ECCV 2024*, pp. 235–251, 2024. doi: 10.1007/978-3-031-73209-6_14. URL https://doi.org/10.1007/978-3-031-73209-6_14.

Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7280–7290, 2023. doi: 10.1109/ICCV51070.2023.00669. URL https://doi.org/10.1109/ICCV51070.2023.00669.

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, 2025a. URL https://doi.org/10.48550/arXiv.2411.04746. Proceedings of the 42nd International Conference on Machine Learning.

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, 2024. URL https://doi.org/10.48550/arXiv.2401.07519.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance, 2025b. URL https://doi.org/10.48550/arXiv.2406.07209. International Conference on Learning Representations (ICLR).

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. doi: 10.1109/CVPR46437.2021.00905. URL https://doi.org/10.1109/CVPR46437.2021.00905.

Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. URL https://www.cns.nyu.edu/~lcv/ssim/.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. OmniGen2: Exploration to advanced multimodal generation, 2025. URL https://arxiv.org/abs/2506.18871v1.

Yi Wu, Ziqiang Li, Heliang Zheng, Chaoyue Wang, and Bin Li. Infinite-id: Identity-preserved personalization via id-semantics decoupling paradigm. In *Computer Vision – ECCV 2024*, pp. 279–296, 2024. doi: 10.1007/978-3-031-73242-3_16. URL https://doi.org/10.1007/978-3-031-73242-3_16.

Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 133:1175–1194, 2025. doi: 10.1007/s11263-024-02227-z. URL https://doi.org/10.1007/s11263-024-02227-z.

Hengyuan Xu, Wei Cheng, Peng Xing, Yixiao Fang, Shuhan Wu, Rui Wang, Xianfang Zeng, Daxin Jiang, Gang Yu, Xingjun Ma, and Yu-Gang Jiang. Withanyone: Towards controllable and id consistent image generation, 2026. URL https://doi.org/10.48550/arXiv.2510.14975. International Conference on Learning Representations (ICLR).

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. Tfg: Unified training-free guidance for diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0704. URL https://doi.org/10.52202/079017-0704.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. URL https://doi.org/10.48550/arXiv.2308.06721.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model, 2023. URL https://doi.org/10.48550/arXiv.2303.09833.

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*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Yimeng Zhang, Tiancheng Zhi, Jing Liu, Shen Sang, Liming Jiang, Qing Yan, Sijia Liu, and Linjie Luo. Id-patch: Robust id association for group photo personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2986–2996, 2025. doi: 10.1109/CVPR52734.2025.00284. URL https://doi.org/10.1109/CVPR52734.2025.00284.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8568–8577, 2023. doi: 10.1109/CVPR52729.2023.00828. URL https://doi.org/10.1109/CVPR52729.2023.00828.

Shangchen Zhou, Kelvin C. K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30599–30611, 2022. doi: 10.52202/068431-2218. URL https://doi.org/10.52202/068431-2218.

Chenyang Zhu, Kai Li, Yue Ma, Chunming He, and Xiu Li. Multibooth: Towards generating all your concepts in an image from text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10923–10931, 2025. doi: 10.1609/aaai.v39i10.33187. URL https://doi.org/10.1609/aaai.v39i10.33187.

A Additional Experimental Details

A.1 Datasets, Metrics, and Implementation

For the fixed-candidate comparison, we construct 707 source–target records from the single- and multi-person subsets of MultiID-Bench (Xu et al., 2026). After three result-blind exclusions caused by invalid candidate localization, 704 records remain: 64 calibration cases and 640 test cases. The split is performed at the source-image level, with 30 calibration source groups and 402 test source groups; the test set contains 195, 366, 39, and 40 target cases from one-, two-, three-, and four-person tasks. Common edited candidates are generated with the FLUX.1-dev-based WithAnyone model and their pixels, target references, and evaluation geometry are frozen before comparing methods. Calibration cases are used for implementation and hyperparameter choices; the 640 test cases are evaluated without further tuning. An unsuccessful method output is replaced by the unchanged candidate and retained in the denominator.

For cross-source/backbone evaluation, we use all 227 multi-person tasks (194 two-person and 33 three-/four-person tasks) with 501 identity references. Each candidate source attempts one image per task without best-of-N selection. Faces are detected in each candidate and Hungarian matching assigns references to candidate faces; the assignment and five-point target geometry are fixed before refinement. Every target is refined independently from its original candidate. Reference–face coverage is measured against all 501 references, while refinement completion is measured against matched targets. The controlled ablations use the same 494 matched targets from the FLUX.2 source, and the Full configuration reuses the corresponding cross-backbone outputs.

Target identity uses cosine similarity to the reference with ArcFace (Deng et al., 2019), FaceNet (Schroff et al., 2015), and AdaFace (Kim et al., 2022). ArcFace is also the guidance recognizer; FaceNet and AdaFace provide complementary evaluation. The fixed-candidate comparison uses ArcFace Glint-R100, while subsequent experiments use W600K-R50, together with FaceNet pretrained on VGGFace2 and AdaFace IR50; absolute ArcFace values are therefore compared only within a protocol. Fixed-candidate preservation reports non-target retention averaged over the three recognizers, normalized RGB ℓ_1, and SSIM (Wang et al., 2004) outside the frozen target box enlarged by $1.50\times$. Later experiments use non-target AdaFace retention and AlexNet LPIPS (Zhang et al., 2018) outside a $1.70\times$ target box. Source- or case-clustered bootstrap intervals use 5,000 resamples for comparison experiments and 10,000 for full ablations.

All experiments use PyTorch, 512×512 images, and an NVIDIA A100 80 GB GPU; all generation, decoding, and recognition networks remain frozen. FLUX.2 and Qwen refiners both use four steps. For FLUX.2 controlled ablations, $\rho_{\text{edit}} = 0.55$, $\rho_{\text{keep}} = 0.02$, and the initial identity-to-context RMS ratio is 2; the corresponding Qwen values are 0.25, 0.02, and 1. Core and halo box scales are 1.20 and 1.70, background pull strength is 0.5, and candidate-trajectory noise decays with the flow schedule. The fixed-candidate refiner uses seed 42 with a stable target-key policy, common WithAnyone candidates use seed 1234, and cross-backbone generation/refinement and controlled ablations use seed 0.

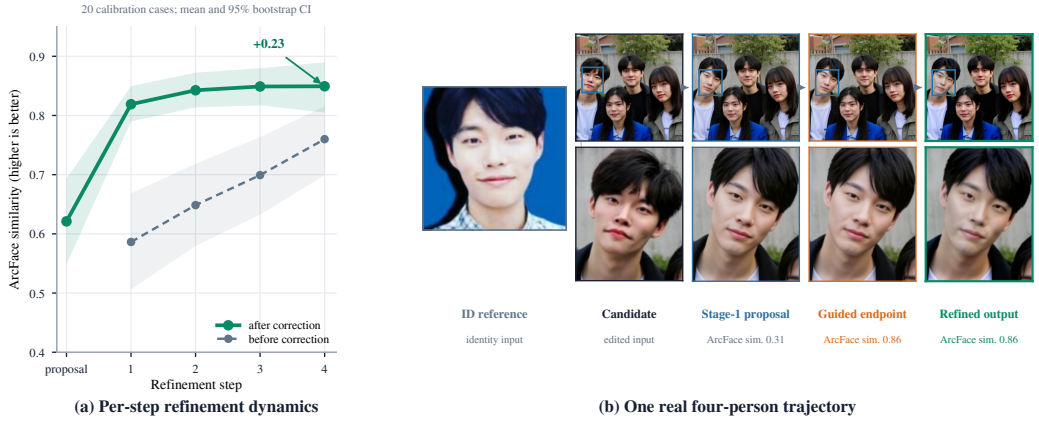
B Additional Comparison Diagnostics

Fig. 5(a) analyzes 20 calibration trajectories. Mean corrected ArcFace similarity rises from about 0.62 at the proposal to 0.85 after four steps (roughly $+0.23$), and corrected endpoint scores remain above pre-correction scores throughout the trajectory. In the four-person example, the guided endpoint reaches 0.86 from a proposal score of 0.31, with the displayed score retained after recomposition while the group composition remains visually consistent.

C Full Ablation Results

C.1 Support-Aware Gradient Calibration

Fig. 4 uses 1,976 step records from 494 SAGC trajectories. The measured global-to-support normalization ratio varies systematically with effective support, consistent with the $\sqrt{q/p}$ relation in Appendix G.3. Counterfactual global calibration delivers $5.63\times$ the intended local guidance in the



(a) Per-step refinement dynamics **(b) One real four-person trajectory**

Figure 5: Four-step refinement dynamics. (a) ArcFace similarity before and after clean-endpoint correction on 20 calibration trajectories; shading denotes 95% bootstrap intervals. (b) A four-person example with the selected face enlarged.

Table 5: Full calibration ablation on 494 paired targets. Win rates denote strictly positive identity gain; NT is non-target AdaFace retention.

Calibration	ΔAdaFace	Ada win (%)	ΔFaceNet	FN win (%)	Outside LPIPS	NT	ΔArcFace
Global RMS	**+0.3655**	99.60	+0.1139	81.58	0.01472	0.9724	**+0.3777**
SAGC (Ours)	+0.3056	**100.00**	**+0.1314**	**97.57**	**0.01257**	**0.9738**	+0.3013

Table 6: Calibration ablation stratified by effective latent-support fraction p. Each quartile uses identical targets for both variants.

Support quartile	Calibration	N	ΔAdaFace	Win (%)	ΔFaceNet	Outside LPIPS	NT
Q1: 0.0059–0.0479	Global RMS	124	+0.3833	100.00	+0.1520	0.01450	0.9613
	SAGC	124	+0.3054	100.00	+0.1689	0.01193	0.9624
Q2: 0.0479–0.0747	Global RMS	124	+0.4062	99.19	+0.1152	0.01402	0.9731
	SAGC	124	+0.3319	100.00	+0.1343	0.01172	0.9742
Q3: 0.0747–0.1182	Global RMS	123	+0.3737	100.00	+0.1093	0.01482	0.9758
	SAGC	123	+0.3109	100.00	+0.1251	0.01291	0.9771
Q4: 0.1182–0.4727	Global RMS	123	+0.2984	99.19	+0.0789	0.01553	0.9796
	SAGC	123	+0.2739	100.00	+0.0970	0.01373	0.9814

smallest face-size group versus $2.37\times$ in the largest. Fig. 6 complements this diagnostic with frozen real-output comparisons across representative face scales.

Across the 494 final outputs (see Tab. 5), SAGC minus global RMS gives FaceNet-gain difference $+0.017457$ (paired 95% CI $[+0.007686, +0.027213]$), outside-LPIPS difference -0.002147 $[-0.002314, -0.001985]$, and non-target-retention difference $+0.001325$ $[+0.000800, +0.001839]$; global RMS retains the larger AdaFace and guidance-ArcFace gains. A separate 780-state matched-gradient diagnostic spans support fractions 1.56%–47.27% and Global/SAGC coefficient ratios 1.45–8.00 (median 3.56).

SAGC improves FaceNet gain and outside LPIPS in every support quartile, while global RMS retains the larger AdaFace gain, as seen in Tab. 6. Together with the paired final-output intervals above, this connects support-aware denominator calibration to the empirical transfer–preservation trade-off at the reported operating point.

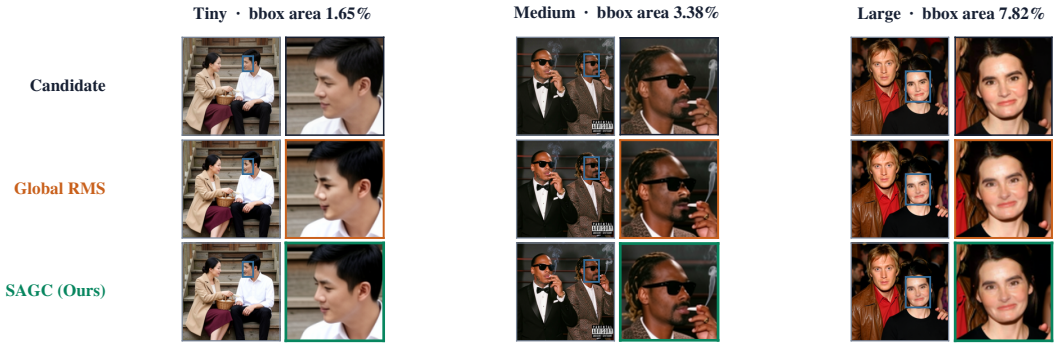


Figure 6: Qualitative comparison of gradient calibration across face scales. For representative tiny, medium, and large target faces, each column shows the frozen candidate followed by refinement with whole-latent global RMS calibration and SAGC. The target location and candidate-defined crop are fixed across the two calibration rules.

Table 7: Full candidate-alignment ablation. Part (a) uses 494 targets; part (b) compares quality on 493 common successful targets, while dynamic alignment completes 493/494 attempts.

Configuration	Complete (%)	ΔAdaFace	Ada win (%)	ΔFaceNet	FN win (%)	Outside LPIPS	NT	ΔArcFace
(a) Initialization: $N = 494$								
Native restart	100.00	+0.2447	97.37	+0.0885	77.13	0.01967	0.9640	+0.2462
Candidate-aligned (Ours)	100.00	+0.3056	100.00	+0.1314	97.57	0.01257	0.9738	+0.3013
(b) Identity coordinates: common $N = 493$								
Bounding-box crop	100.00	+0.2193	98.78	+0.1153	92.70	0.01256	0.9737	+0.2086
Dynamic five-point	99.80	+0.2722	99.39	+0.1174	96.15	0.01254	0.9719	+0.2662
Fixed five-point (Ours)	100.00	+0.3058	100.00	+0.1313	97.57	0.01257	0.9738	+0.3015

Table 8: Spatial-policy ablation on 494 targets. All configurations complete every target.

Spatial policy	ΔAdaFace	Ada win (%)	ΔFaceNet	FN win (%)	Outside LPIPS	NT	ΔArcFace
Uniform	+0.0763	98.99	+0.0356	87.85	0.00659	0.9589	+0.0748
Core only	+0.2654	100.00	+0.1152	96.96	0.03256	0.9453	+0.2610
Core + pull	+0.3064	100.00	+0.1314	96.96	0.01196	0.9757	+0.3020

Table 9: Calibration sweep for the task-matched training-free guidance comparison. Stars mark the frozen baseline choices; Ours retains its canonical configuration.

Method	Strength	ΔAdaFace	ΔFaceNet	LPIPS_{out}	Success
State guidance (adapt.)	0.25	0.0107	-0.0030	0.0149	53/53
State guidance (adapt.)	0.5	0.0317	0.0067	0.0133	53/53
State guidance (adapt.)	1⋆	0.0822	0.0330	0.0113	53/53
State guidance (adapt.)	2	0.2302	0.0719	0.0304	53/53
State guidance (adapt.)	4	0.0187	-0.2732	0.1722	53/53
TFG mixed (adapt.)	0.25	0.0521	0.0167	0.0151	53/53
TFG mixed (adapt.)	0.5	0.1004	0.0429	0.0135	53/53
TFG mixed (adapt.)	1	0.1761	0.0800	0.0115	53/53
TFG mixed (adapt.)	2⋆	0.3211	0.1279	0.0133	53/53
TFG mixed (adapt.)	4	0.0356	-0.2674	0.1668	53/53
Ours (canonical)	–	0.2860	0.1264	0.0119	53/53

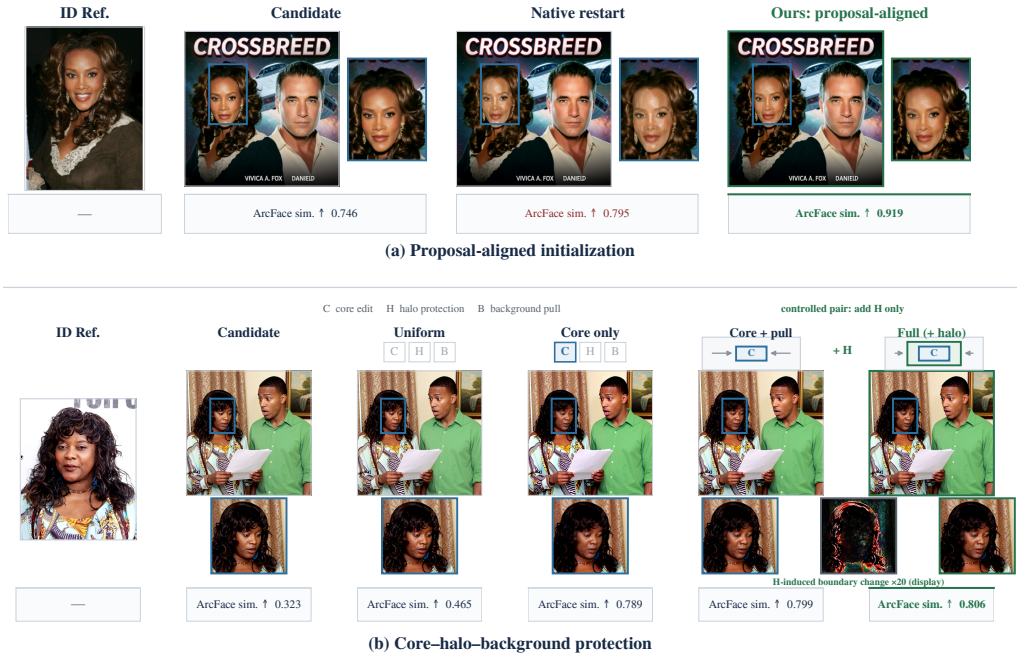


Figure 7: Qualitative ablation of candidate-aligned initialization and spatial protection. The 20× residual inset visualizes the localized effect of halo protection.

C.2 Candidate Alignment and Spatial Policy

Candidate-aligned initialization improves all reported quality metrics over native restart. Fixed five-point alignment gives the strongest identity gain and avoids the one failure of dynamic redetection (see Tab. 7). Spatially, concentrating the larger initialization budget in the core raises identity gain but increases drift; background pull restores preservation, and halo protection slightly relaxes the pull around facial boundaries (see Tab. 8). In the displayed halo example (Fig. 7), Core + pull and Full achieve ArcFace scores of 0.799 and 0.806, respectively.

D Controlled Guidance, Transport, and Boundary Diagnostics

D.1 Training-Free Guidance Comparison

We compare the canonical endpoint update with task-matched state-gradient guidance and mixed TFG guidance on the same FLUX.2-KV candidates, as seen in Fig. 8. All three arms share the frozen backbone, candidate/reference conditioning, ArcFace and source-preservation objectives, fixed face geometry, candidate-aligned initialization, SAGC, Core–Halo–Background controls, seed, and four-step schedule. The comparison therefore changes the guidance update while preserving the identity-refinement task and spatial policy. The state arm propagates the decoded-clean objective through the frozen flow-model Jacobian; the mixed TFG arm combines state and clean branches with equal weights. Ours applies the accepted clean-endpoint correction and exact recomposition.

The source-image split contains 24 calibration images with 53 targets and 203 held-out images with 441 targets. As seen in Tab. 9, state and mixed-TFG strengths are selected by maximizing mean AdaFace gain subject to outside LPIPS ≤ 0.020; the tested grid is $\{0.25, 0.5, 1, 2, 4\}$. Ours retains the canonical setting from the main experiments.

Compared with state guidance (see Tab. 10 and 11), Ours increases mean AdaFace and FaceNet gains by 0.2193 and 0.0949, with paired 95% CIs $[0.2079, 0.2302]$ and $[0.0870, 0.1035]$; outside LPIPS increases by 0.000526 $[0.000254, 0.000708]$. Relative to mixed TFG, Ours increases FaceNet gain by 0.0330 $[0.0212, 0.0457]$; the Ours-minus-TFG outside-LPIPS difference is -0.003982



Figure 8: Qualitative comparison of training-free guidance controls on three held-out targets. Each row shows the identity reference, frozen candidate, state guidance, mixed TFG, and Ours.

Table 10: Held-out training-free guidance comparison on 441 targets from 203 images. FN+ is the fraction with positive FaceNet gain; NT det. is non-target detection retention.

Method	ΔAdaFace	ΔFaceNet	FN+	LPIPS_{out}	NT-ID	NT det.
State guidance (adapt.)	0.0885	0.0370	72.8%	0.0121	0.9722	99.5%
TFG mixed (adapt.)	0.2962	0.0989	83.4%	0.0166	0.9718	99.5%
Ours	0.3078	0.1319	97.5%	0.0127	0.9739	99.5%

$[-0.005755, -0.002422]$. The AdaFace difference is $+0.0117$ with CI $[-0.0004, 0.0244]$. Confidence intervals use 10,000 paired image-cluster bootstrap draws with targets weighted equally.

The endpoint update still differentiates through the frozen decoder and external identity/preservation objectives. The compute difference above specifically reflects the absence of a backward pass through the flow backbone.

Table 11: Paired held-out differences for Ours minus each task-matched guidance control.

Comparator	Metric	Difference	95% CI
State guidance (adapt.)	ΔAdaFace	+0.219304	[+0.207854,+0.230246]
State guidance (adapt.)	ΔFaceNet	+0.094874	[+0.087042,+0.103466]
State guidance (adapt.)	Outside LPIPS	+0.000526	[+0.000254,+0.000708]
TFG mixed (adapt.)	ΔAdaFace	+0.011658	[-0.000420,+0.024403]
TFG mixed (adapt.)	ΔFaceNet	+0.032991	[+0.021211,+0.045685]
TFG mixed (adapt.)	Outside LPIPS	-0.003982	[-0.005755,-0.002422]

Table 12: Compute profile for the held-out guidance comparison. Core time is synchronized profiling on three fixed calibration targets; recorded time is the instrumented held-out call time; maximum memory is peak allocated CUDA memory.

Method	Backbone predictions	Backbone backward	Recorded s/target	Core s/target	Max. GiB
State guidance (adapt.)	4	4	6.386	4.409	77.19
TFG mixed (adapt.)	4	4	6.312	4.383	77.20
Ours	4	0	4.697	2.681	38.75

Table 13: Complete calibration sweep for the clean-correction transport ablation. Stars mark the independently selected settings.

Rule	Strength	ΔAdaFace	ΔFaceNet	LPIPS_{out}	Success
Endpoint	0.25	0.0972	0.0387	0.0154	53/53
Endpoint	0.5	0.1810	0.0793	0.0139	53/53
Endpoint	1	0.2853	0.1267	0.0119	53/53
Endpoint*	2	0.3624	0.1513	0.0104	53/53
Endpoint	4	0.2658	0.0930	0.0129	53/53
Clean/velocity	0.25	0.0707	0.0237	0.0159	53/53
Clean/velocity	0.5	0.1441	0.0599	0.0147	53/53
Clean/velocity	1	0.2384	0.1014	0.0130	53/53
Clean/velocity*	2	0.2966	0.1171	0.0124	53/53
Clean/velocity	4	0.2424	0.0892	0.0134	53/53
Direct injection	0.25	0.1243	0.0529	0.0149	53/53
Direct injection	0.5	0.2241	0.0993	0.0131	53/53
Direct injection	1	0.3636	0.1484	0.0111	53/53
Direct injection*	2	0.4102	0.1331	0.0102	53/53
Direct injection	4	0.2637	0.0778	0.0129	53/53
Endpoint, no acceptance	0.25	0.0974	0.0386	0.0154	53/53
Endpoint, no acceptance	0.5	0.1808	0.0788	0.0139	53/53
Endpoint, no acceptance	1	0.2854	0.1258	0.0119	53/53
Endpoint, no acceptance*	2	0.3783	0.1497	0.0102	53/53
Endpoint, no acceptance	4	0.2050	-0.0575	0.0854	53/53

D.2 Controlled Clean-Correction Transport

We next keep the candidate/reference inputs, initialization, ArcFace and preservation objectives, SAGC, spatial policy, clipping, four backbone predictions, and eight external gradient calls fixed, and vary how an accepted clean correction is propagated to the next state. The three primary rules correspond to Eqs. 25–27; a fourth arm removes the acceptance gate (see Fig. 9). Each arm independently selects strength on the same 53-target calibration set using the same five-point grid and AdaFace/LPIPS selection rule, as seen in Tab. 13. All four select $s = 2$ before evaluation on the 441-target held-out set. The predeclared canonical endpoint setting $s = 1$ is retained as an additional reference.

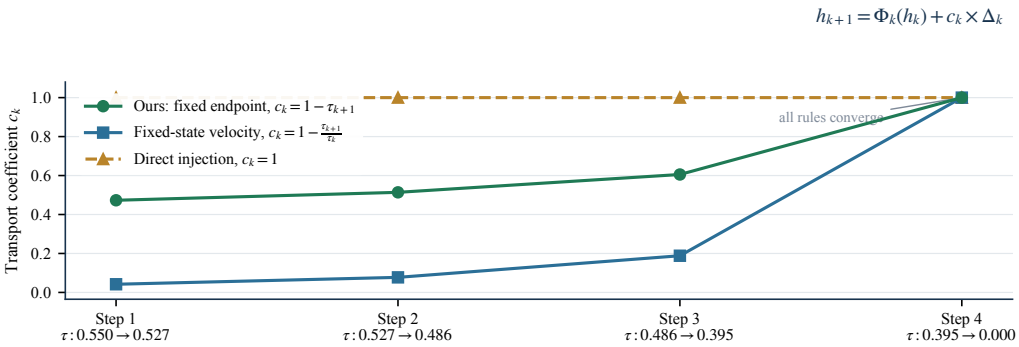


Figure 9: Step-dependent transport coefficients for propagating the same accepted clean correction. Writing the next state as $\mathbf{h}_{k+1} = \Phi_k(\mathbf{h}_k) + c_k\Delta_k$, fixed-complementary-endpoint recomposition gives $c_k = 1 - \tau_{k+1}$, fixed-current-state clean/velocity conversion gives $c_k = 1 - \tau_{k+1}/\tau_k$, and direct injection gives $c_k = 1$. The coefficients follow the four-step schedule used in the controlled transport experiment and converge at the final step.

Table 14: Held-out transport results on 441 targets. All arms complete all targets.

Rule	s	ΔAdaFace	ΔFaceNet	FN+	LPIPS$_{out}$	NT-ID	NT det.
Endpoint	2	0.3753	0.1504	93.7%	0.0115	0.9741	439/441
Clean/velocity	2	0.3005	0.1162	88.9%	0.0131	0.9730	439/441
Direct injection	2	0.3952	0.1344	84.6%	0.0111	0.9735	440/441
Endpoint, no acceptance	2	0.3796	0.1441	88.0%	0.0110	0.9743	439/441
Endpoint (canonical)	1	0.3081	0.1320	97.7%	0.0127	0.9739	439/441

Table 15: Paired held-out differences for the transport ablation. Differences are endpoint anchor minus comparator.

Endpoint anchor	Comparator	ΔAdaFace difference [95% CI]	ΔFaceNet difference [95% CI]	LPIPS$_{out}$ difference [95% CI]
Tuned	Clean/velocity	+0.0748 [+0.0687,+0.0810]	+0.0342 [+0.0289,+0.0397]	-0.0017 [-0.0018,-0.0015]
Tuned	Direct injection	-0.0199 [-0.0273,-0.0125]	+0.0160 [+0.0083,+0.0239]	+0.0004 [+0.0003,+0.0005]
Tuned	Endpoint, no acceptance	-0.0043 [-0.0093,+0.0007]	+0.0064 [+0.0015,+0.0113]	+0.0005 [+0.0004,+0.0007]
Tuned	Endpoint (canonical)	+0.0672 [+0.0593,+0.0753]	+0.0184 [+0.0128,+0.0241]	-0.0012 [-0.0013,-0.0011]
Canonical	Clean/velocity	+0.0076 [-0.0003,+0.0154]	+0.0158 [+0.0109,+0.0206]	-0.0005 [-0.0006,-0.0003]
Canonical	Direct injection	-0.0871 [-0.0997,-0.0748]	-0.0024 [-0.0130,+0.0083]	+0.0016 [+0.0015,+0.0017]
Canonical	Endpoint, no acceptance	-0.0715 [-0.0816,-0.0617]	-0.0121 [-0.0200,-0.0042]	+0.0017 [+0.0016,+0.0018]

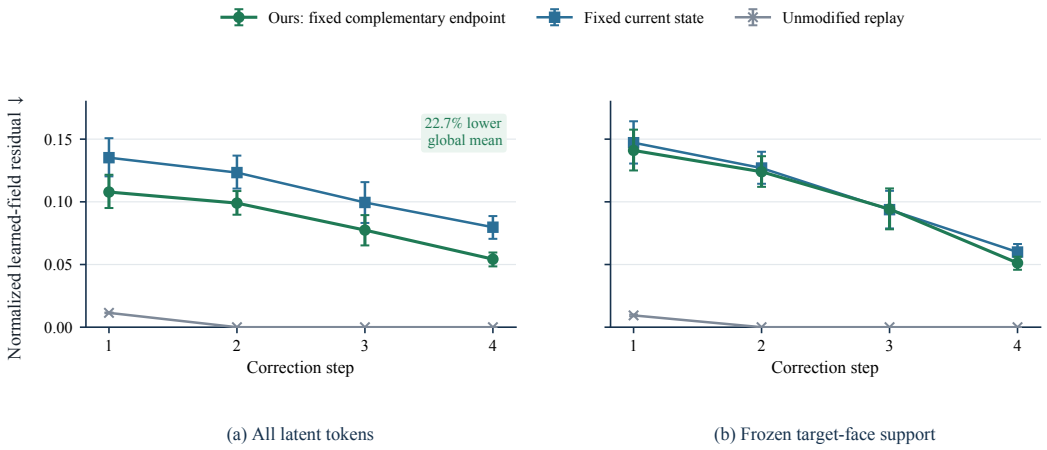
At the symmetrically selected settings (see Tab. 14 and 15), endpoint recomposition improves over clean/velocity conversion by $+0.0748$ AdaFace $[0.0687, 0.0810]$, $+0.0342$ FaceNet $[0.0289, 0.0397]$, and -0.0017 outside LPIPS $[-0.0018, -0.0015]$. The three transport rules exhibit different identity–preservation trade-offs, confirming that correction transport affects refinement behavior beyond the clean correction itself. The canonical endpoint setting also exceeds the selected clean/velocity control on FaceNet by $+0.0158$ $[0.0109, 0.0206]$ with 0.0005 lower LPIPS; its AdaFace difference $+0.0076$ has interval $[-0.0003, 0.0154]$.

At $s = 4$ on calibration, the guarded endpoint arm retains FaceNet gain $+0.09305$ with outside LPIPS 0.01288, while the unguarded arm reaches -0.05751 and 0.08541. Across 1,764 held-out steps at $s = 2$, the unguarded arm admits 275 clean corrections that increase the current identity loss; accepted endpoint updates admit zero such trials (Tab. 16).

Local learned-field diagnostic. As shown in Fig. 10, using fresh backbone predictions on 32 input-selected images, the same accepted endpoint correction yields normalized global residual 0.08464 under the fixed-complementary-endpoint constraint and 0.10945 under the fixed-

Table 16: Realized acceptance budget on the 441-target transport evaluation. ID viol. is the fraction of clean trials admitted with increased identity loss.

Rule	s	Trial evals.	First	Half	Reject	ID viol.
Endpoint	2	5.34	66.6%	31.0%	2.5%	0.0%
Clean/velocity	2	5.76	56.1%	43.4%	0.5%	0.0%
Direct injection	2	5.75	56.3%	39.6%	4.1%	0.0%
Endpoint, no acceptance	2	4.00	100.0%	0.0%	0.0%	15.6%
Endpoint (canonical)	1	4.07	98.2%	1.6%	0.1%	0.0%



(a) All latent tokens

(b) Frozen target-face support

Figure 10: Learned-field residual under matched clean corrections on 32 input-selected images. Curves report the normalized residual across the four correction steps for fixed-complementary-endpoint recomposition, the fixed-current-state constraint, and unmodified-state replay, measured over (a) all latent tokens and (b) the frozen target-face support. Error bars are 95% image-cluster bootstrap intervals. Fixed-complementary-endpoint recomposition has a 22.7% lower global mean residual than the fixed-current-state constraint.

current-state constraint, a 22.7% reduction; the paired difference is -0.02481 with 95% CI $[-0.02836, -0.02145]$. On target-face support the corresponding means are 0.1026 and 0.1070, and the paired difference is -0.00435 $[-0.00717, -0.00113]$. Unmodified-state replay gives normalized residual 0.00286 globally and approximately 0.0024 on face support. These measurements characterize local field agreement separately from the exact recomposition identities in Appendix G.2.

D.3 Boundary Protection from Candidate-Trajectory Pull

The halo mechanism is evaluated at matched pre-pull states to isolate the spatial effect of the pull operator. We use 195 candidate images outside the earlier 32-image diagnostic, with one input-selected target per image and four canonical states per target. The visual results are illustrated in Fig. 11. The 780 matched states yield 3,120 nonzero-pull measurements across the reported intervention branches. Each image contributes the mean of its four states, giving 195 independent image units. Let $\mathbf{z}_k$ denote the shared pre-pull recomposed latent and $\mathbf{q}_k$ the shared candidate-trajectory anchor. The halo and no-halo pull maps use the canonical coefficient $\lambda = 0.5$:

$$P_H = \lambda(1 - M_H), \qquad P_{NH} = \lambda(1 - M_C), \tag{16}$$

where M_C and M_H are the core and halo masks. Boundary Pull Interference (BPI) measures the RGB RMS change caused by the pull in a fixed band B_r around the candidate-fixed core boundary,

$$\mathrm{BPI}_r(m) = \sqrt{\frac{1}{3|B_r|} \sum_{p \in B_r} \|I_k^m(p) - I_k^0(p)\|_2^2}, \qquad r = 4 \text{ pixels}, \tag{17}$$

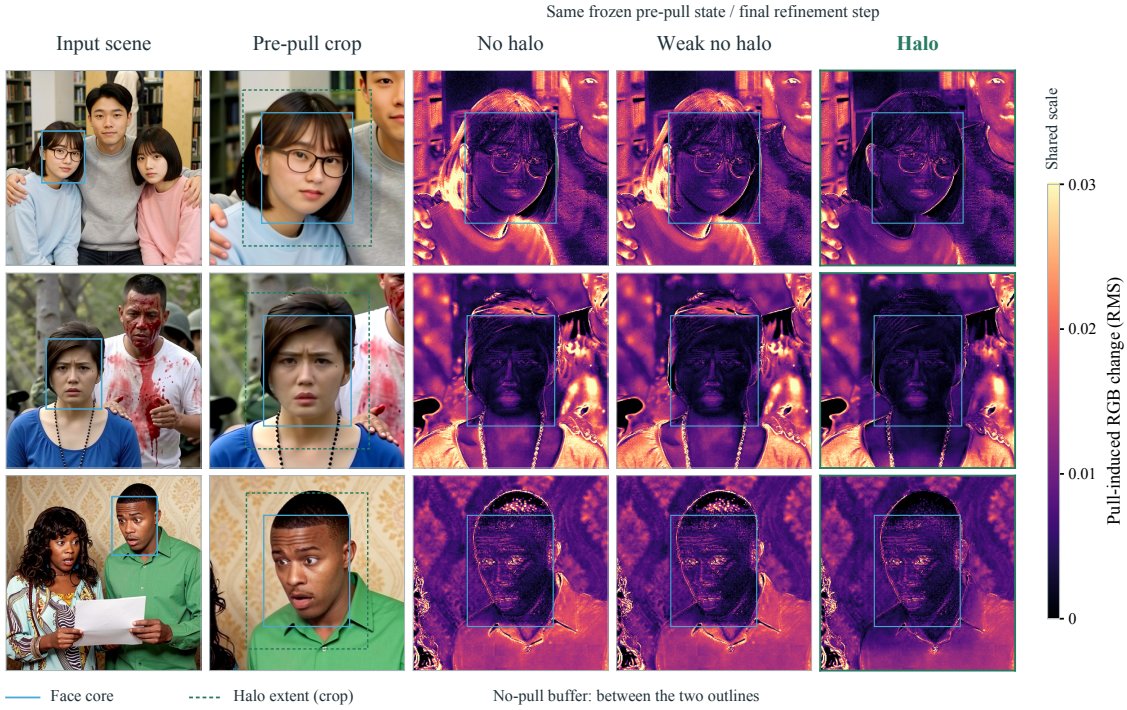


Figure 11: Spatial visualization of the matched-state halo intervention. The first two columns show the input scene and candidate-defined target crop. The heatmaps show pixel-space RMS changes induced by candidate-trajectory pull under the same frozen pre-pull state and shared intervention scale. The blue and green contours denote the face core and halo extent, respectively; the region between them is the no-pull buffer. Compared with no-halo and energy-matched weak no-halo controls, halo protection suppresses pull-induced changes near the editable face boundary while retaining farther-background intervention.

Table 17: Matched-state halo intervention on 195 images. Lower values indicate less pull-induced boundary disturbance.

Measure	No halo	Weak no halo	Halo
BPI, $r = 4$	0.010645	0.009969	0.004810
BPGJ, $h = 4$	0.009249	0.008697	0.005301
CB-LPIPS, $r = 16$	0.001963	0.001740	0.000809
PLR, $r = 4$	0.716291	0.716701	0.326788

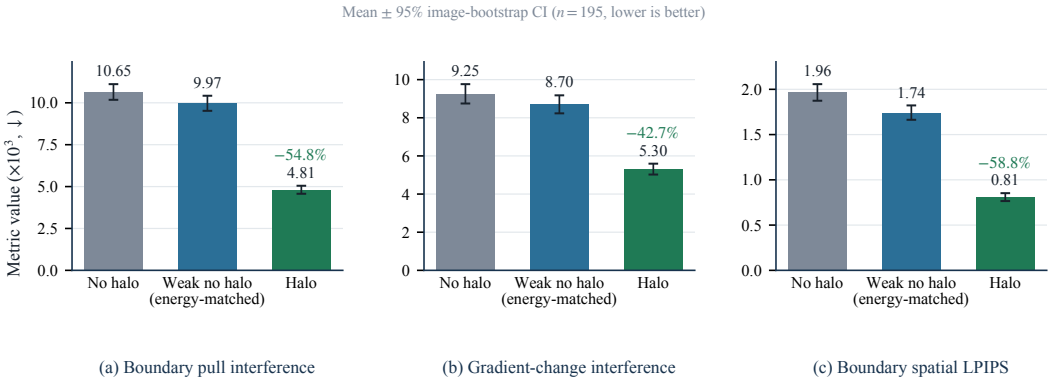
with $I_k^0 = \mathcal{V}(\mathbf{z}_k)$ and I_k^m the decoded post-pull state. BPGJ measures the pull-induced change in normal-gradient structure along the core boundary, CB-LPIPS averages spatial AlexNet LPIPS between I_k^m and I_k^0 over a fixed boundary band, and PLR is primary-band BPI divided by the RGB RMS intervention on the common background (plus 10^{-6}).

We also construct a weak no-halo control with coefficient

$$\lambda_w = \lambda \frac{\|(1 - M_H)(\mathbf{q}_k - \mathbf{z}_k)\|_2}{\|(1 - M_C)(\mathbf{q}_k - \mathbf{z}_k)\|_2}, \tag{18}$$

which matches the halo intervention in global latent ℓ_2 norm. The maximum relative matching error is 2.30×10^{-6}.

As shown in Tab. 17 and Fig. 12, Halo reduces primary BPI by 54.8% relative to no halo and 51.8% relative to the energy-matched weak no-halo control. Halo and the original no-halo policy both retain the canonical 0.5 residual contraction on their common active background; the globally



(a) Boundary pull interference (b) Gradient-change interference (c) Boundary spatial LPIPS

Figure 12: Primary matched-state halo intervention controls on 195 images. Bars show mean ± 95% image-bootstrap intervals for boundary pull interference, gradient-change interference, and boundary spatial LPIPS; lower is better. The weak no-halo control matches halo in global latent intervention norm. Relative to no halo, halo reduces the three mean interference measures by 54.8%, 42.7%, and 58.8%, respectively.

Table 18: Paired halo differences with 10,000 image-bootstrap resamples.

Measure	Halo − no halo [95% CI]	Halo − weak no halo [95% CI]
BPI, $r = 4$	-0.005836 [-0.006185,-0.005491]	-0.005160 [-0.005462,-0.004863]
BPGJ, $h = 4$	-0.003948 [-0.004319,-0.003594]	-0.003396 [-0.003715,-0.003095]
CB-LPIPS, $r = 16$	-0.001155 [-0.001235,-0.001075]	-0.000931 [-0.000993,-0.000873]
PLR, $r = 4$	-0.389503 [-0.412886,-0.367290]	-0.389913 [-0.413225,-0.367736]

Table 19: Boundary Pull Interference by input face-area quartile.

Quartile	Images	No-halo BPI	Halo BPI	Halo − no halo [95% CI]	Reduction vs weak no halo
Q1	48	0.011254	0.005902	-0.005352 [-0.006083,-0.004665]	46.2%
Q2	49	0.011626	0.005497	-0.006128 [-0.006931,-0.005351]	50.5%
Q3	49	0.010156	0.004541	-0.005615 [-0.006172,-0.005070]	52.6%
Q4	49	0.009558	0.003321	-0.006237 [-0.006905,-0.005557]	59.8%

Table 20: Predefined boundary-measure sensitivities.

Measure / setting	Halo − no halo [95% CI]	Halo − weak no halo [95% CI]
BPI $r = 4$	-0.005836 [-0.006185,-0.005491]	-0.005160 [-0.005462,-0.004863]
BPI $r = 8$	-0.005998 [-0.006334,-0.005670]	-0.005317 [-0.005607,-0.005034]
BPI $r = 16$	-0.006047 [-0.006357,-0.005734]	-0.005366 [-0.005639,-0.005101]
BPI $r = 24$	-0.005779 [-0.006086,-0.005474]	-0.005083 [-0.005342,-0.004828]
BPGJ $h = 2$	-0.002539 [-0.002847,-0.002247]	-0.002128 [-0.002392,-0.001880]
BPGJ $h = 4$	-0.003948 [-0.004319,-0.003594]	-0.003396 [-0.003715,-0.003095]
BPGJ $h = 8$	-0.005776 [-0.006195,-0.005373]	-0.005073 [-0.005436,-0.004725]
CB-LPIPS $r = 8$	-0.001181 [-0.001263,-0.001100]	-0.000956 [-0.001018,-0.000896]
CB-LPIPS $r = 16$	-0.001155 [-0.001235,-0.001075]	-0.000931 [-0.000993,-0.000873]
CB-LPIPS $r = 24$	-0.001114 [-0.001194,-0.001037]	-0.000892 [-0.000952,-0.000835]
Final-state BPI, pull 0.25	-0.002042 [-0.002244,-0.001848]	–
Final-state BPI, pull 0.50	-0.004731 [-0.005153,-0.004331]	-0.004148 [-0.004538,-0.003779]
Final-state BPI, pull 0.75	-0.008160 [-0.008831,-0.007533]	–

Table 21: Candidate-stage coverage and refinement completion. Coverage is measured against all 501 references and completion against matched targets.

Candidate source	Generated images	Matched targets N	Coverage (%)	Candidate AdaFace	FLUX.2 complete (%)	Qwen complete (%)
FLUX.2	227/227	494	98.60	0.2935	100.00	100.00
Qwen-Image-Edit	227/227	477	95.21	0.3845	100.00	100.00
OmniGen2	227/227	489	97.60	0.1742	100.00	100.00
WithAnyone	220/227	483	96.41	0.5348	100.00	100.00

Table 22: Full identity/preservation matrix. Both refiners complete all matched targets. NT uses 494, 475, 488, and 483 applicable targets for the four sources; other metrics use N.

Candidate source	Refiner	ΔAdaFace	Win (%)	ΔFaceNet	ΔArcFace	Outside LPIPS	NT	N
FLUX.2	FLUX.2	+0.3056	100.00	+0.1314	+0.3013	0.01257	0.9738	494
	Qwen	+0.1335	83.20	+0.0493	+0.1295	0.02995	0.8925	494
Qwen-Image-Edit	FLUX.2	+0.2695	100.00	+0.1121	+0.2651	0.01294	0.9800	477
	Qwen	+0.1479	94.97	+0.0591	+0.1463	0.01290	0.9575	477
OmniGen2	FLUX.2	+0.3521	99.80	+0.1759	+0.3563	0.01334	0.9748	489
	Qwen	+0.1555	88.55	+0.0860	+0.1602	0.02622	0.9332	489
WithAnyone	FLUX.2	+0.1623	99.79	+0.0641	+0.1631	0.01361	0.9835	483
	Qwen	+0.0450	74.95	+0.0220	+0.0438	0.02433	0.9531	483

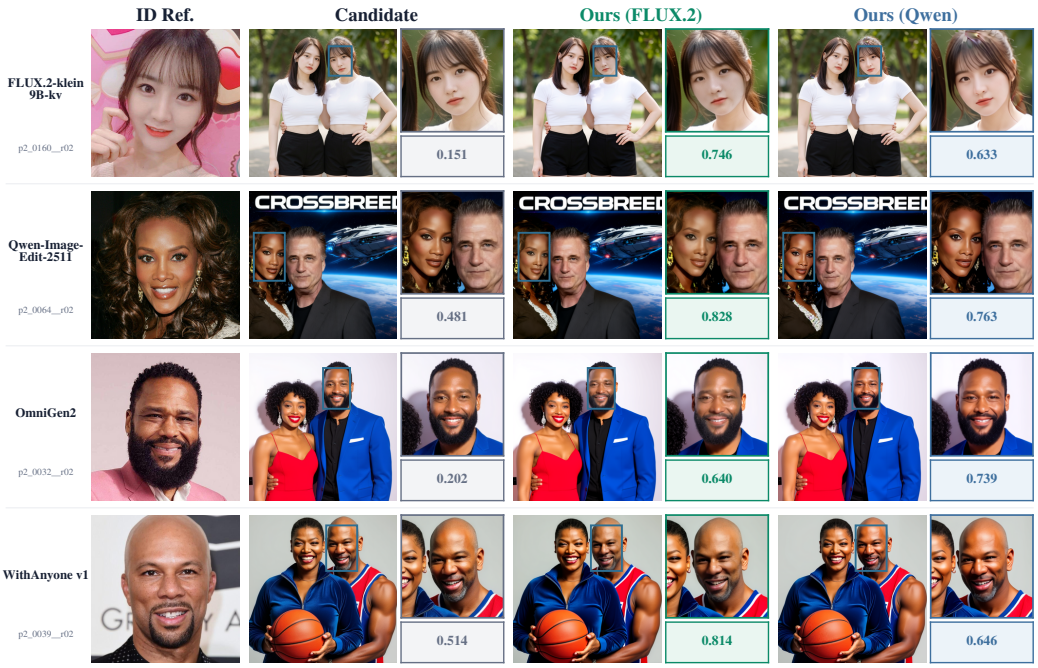


Figure 13: Cross-backbone refinement of candidates from four upstream generators. Both refiners receive the same candidate, target reference, and frozen assignment in each row; footer values are ArcFace similarities.

weakened control has mean residual ratio 0.5310. The effect is consistent across the four input face-area quartiles, with halo-minus-no-halo BPI differences of −0.005352, −0.006128, −0.005615, and −0.006237, respectively.

The original 494-target final-output ablation remains the operating-point comparison: Core + pull and Full (+ halo) have FaceNet gains 0.13136 and 0.13140, respectively; Full minus no-halo outside LPIPS is +0.000606 [+0.000539, +0.000676] and non-target retention is −0.001935 [−0.002611, −0.001366]. The matched-state measurements above isolate the complementary mech-

Table 23: Measured four-step refinement time and peak allocated CUDA memory at 512×512 on an A100. Time excludes candidate generation and model loading.

	FLUX.2 refiner		Qwen refiner	
Candidate source	Time (s)	Memory (GiB)	Time (s)	Memory (GiB)
FLUX.2	7.92	38.75	40.62	54.69
Qwen-Image-Edit	7.55	38.74	40.20	54.69
OmniGen2	7.60	38.75	40.60	54.69
WithAnyone	7.54	38.75	40.20	54.68

anism: halo reduces the direct disturbance created by candidate-trajectory pull at the editable boundary while retaining the farther-background anchoring coefficient.

E Full Cross-Source/Backbone Results and Cost

Both refiners complete every matched target, yielding 3,886 refinement outputs across eight pairings (Tab. 21). WithAnyone has seven candidate-generation failures due to reference-face extraction; these tasks remain included in generation and coverage accounting.

Both backbones improve candidates from all four sources (Tab. 22). The FLUX.2 refiner produces the stronger mean identity gain in each paired comparison. Thus both structurally different backbones support the same post-hoc refinement formulation. Fig. 13 shows that both retain the established person placement, clothing, objects, and neighboring people while improving the selected face.

F Complete Refinement Procedure

For completeness, we summarize the full post-hoc refinement procedure below (see Alg. 1). The notation follows Sec. 3.

G Main Theoretical Properties

G.1 Flow Consistency: Exact Recomposition and Zero-Correction Limit

From Eqs. 5 and 6,

$$\mathbf{a}_k = \mathbf{h}_k - \tau_k \mathbf{d}_k, \tag{19}$$

and

$$\mathbf{b}_k = \mathbf{h}_k + (1 - \tau_k)\mathbf{d}_k. \tag{20}$$

Substituting them into Eq. 4 gives

$$\begin{aligned}(1-\tau_k)\mathbf{a}_k + \tau_k \mathbf{b}_k &= (1-\tau_k)(\mathbf{h}_k - \tau_k \mathbf{d}_k) + \tau_k[\mathbf{h}_k + (1-\tau_k)\mathbf{d}_k] \\ &= [(1-\tau_k) + \tau_k]\mathbf{h}_k - \tau_k(1-\tau_k)\mathbf{d}_k + \tau_k(1-\tau_k)\mathbf{d}_k \\ &= \mathbf{h}_k.\end{aligned} \tag{21}$$

Hence the endpoint decomposition is an exact algebraic reconstruction of the current state.

For the zero-correction case,

$$\bar{\mathbf{a}}_k = \mathbf{a}_k. \tag{22}$$

Equation 11 gives

$$\begin{aligned}\widetilde{\mathbf{h}}_{k+1} &= (1-\tau_{k+1})\mathbf{a}_k + \tau_{k+1}\mathbf{b}_k \\ &= (1-\tau_{k+1})(\mathbf{h}_k - \tau_k \mathbf{d}_k) + \tau_{k+1}[\mathbf{h}_k + (1-\tau_k)\mathbf{d}_k] \\ &= \mathbf{h}_k + (\tau_{k+1} - \tau_k)\mathbf{d}_k.\end{aligned} \tag{23}$$

Equation 23 is exactly the deterministic Euler step for the linear flow parameterization. Thus the recomposition stage recovers the native deterministic Euler update at zero clean correction.

Algorithm 1 Post-hoc Training-Free Identity Refinement

Require: Edited candidate $\mathbf{C}$, identity reference $\mathbf{R}$, frozen encoder $\mathcal{U}$, decoder $\mathcal{V}$, flow backbone $\mathcal{B}_\vartheta$, decreasing flow coordinates $\{\tau_k\}_{k=0}^{K}$
Ensure: Refined image $\mathbf{O}$

$\mathbf{c} \leftarrow \mathcal{U}(\mathbf{C})$
Detect the target face on $\mathbf{C}$ and fix candidate alignment $\Pi_{\mathbf{C}}$
$\mathbf{r}_{\mathrm{id}} \leftarrow \mathcal{Q}(\Pi_{\mathbf{R}}(\mathbf{R}))$
Sample fixed $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
Construct $\rho(\mathbf{u})$ using Eq. 1 and initialize $\mathbf{h}_0$ using Eq. 2
for $k = 0, \ldots, K-1$ **do**
  $\mathbf{d}_k \leftarrow \mathcal{B}_\vartheta(\mathbf{h}_k, \tau_k; \boldsymbol{\psi})$
  $\mathbf{a}_k \leftarrow \mathbf{h}_k - \tau_k \mathbf{d}_k$ ▷ clean endpoint
  $\mathbf{b}_k \leftarrow \mathbf{h}_k + (1-\tau_k)\mathbf{d}_k$ ▷ complementary endpoint
  $\mathbf{Y}_k \leftarrow \mathcal{V}(\mathbf{a}_k)$
  Evaluate $\mathcal{J}_{\mathrm{ctx}}$ and $\mathcal{J}_{\mathrm{id}}$
  $\mathbf{r}_k \leftarrow \nabla_{\mathbf{a}_k}[u_k \mathcal{J}_{\mathrm{ctx}}]$
  $\mathbf{s}_k \leftarrow \nabla_{\mathbf{a}_k} \mathcal{J}_{\mathrm{id}}$
  Compute SAGC scale γ_k
  $\bar{\mathbf{s}}_k \leftarrow \gamma_k \mathbf{s}_k$
  $\mathbf{m}_k \leftarrow \mathbf{r}_k + \bar{\mathbf{s}}_k$
  Apply the global RMS safety cap when required
  Backtrack over trial multipliers and obtain accepted $\bar{\mathbf{a}}_k$ using Eqs. 9–10
  $\widetilde{\mathbf{h}}_{k+1} \leftarrow (1-\tau_{k+1})\bar{\mathbf{a}}_k + \tau_{k+1}\mathbf{b}_k$ ▷ exact recomposition
  Construct candidate trajectory $\mathbf{q}_{k+1}$ using Eq. 13
  Apply Eq. 15 to obtain $\mathbf{h}_{k+1}$
end for
$\mathbf{O} \leftarrow \mathcal{V}(\mathbf{h}_K)$
return $\mathbf{O}$

G.2 Endpoint Transport and Local Learned-Field Agreement

Let

$$\Delta\mathbf{a}_k = \bar{\mathbf{a}}_k - \mathbf{a}_k, \qquad \mathbf{E}_k = \mathbf{h}_k + (\tau_{k+1} - \tau_k)\mathbf{d}_k. \tag{24}$$

Fixing the complementary endpoint and applying Eq. 11 gives

$$\widetilde{\mathbf{h}}_{k+1} = \mathbf{E}_k + (1-\tau_{k+1})\Delta\mathbf{a}_k. \tag{25}$$

The same accepted clean correction can be transported under other constraints. Holding the current state fixed and converting the corrected clean estimate back to a velocity yields, for $\tau_k > 0$,

$$\mathbf{h}_{k+1}^{\mathrm{vel}} = \mathbf{E}_k + \left(1 - \frac{\tau_{k+1}}{\tau_k}\right)\Delta\mathbf{a}_k, \tag{26}$$

whereas unit state injection yields

$$\mathbf{h}_{k+1}^{\mathrm{dir}} = \mathbf{E}_k + \Delta\mathbf{a}_k. \tag{27}$$

For $0 \le \tau_{k+1} \le \tau_k \le 1$, the corresponding coefficients satisfy

$$0 \le 1 - \tau_{k+1}/\tau_k \le 1 - \tau_{k+1} \le 1. \tag{28}$$

Thus the controlled transport ablation in Appendix D.2 changes the temporal propagation of the same accepted clean correction while keeping the remaining optimization pipeline fixed. A direct-state implementation with coefficient $1 - \tau_{k+1}$ is algebraically equivalent to Eq. 25.

The learned-field agreement of the corrected fixed-$\mathbf{b}_k$ segment can be measured at the corresponding same-time state

$$\mathbf{h}'_k = \mathbf{h}_k + (1-\tau_k)\Delta\mathbf{a}_k. \tag{29}$$

Its residual relative to the frozen backbone field is

$$\begin{aligned} \mathbf{r}_k^{\mathrm{field}} &= \mathcal{B}_\vartheta(\mathbf{h}'_k, \tau_k; \boldsymbol{\psi}) - (\mathbf{b}_k - \bar{\mathbf{a}}_k) \\ &= \mathcal{B}_\vartheta(\mathbf{h}_k + (1-\tau_k)\Delta\mathbf{a}_k, \tau_k; \boldsymbol{\psi}) - \mathcal{B}_\vartheta(\mathbf{h}_k, \tau_k; \boldsymbol{\psi}) + \Delta\mathbf{a}_k. \end{aligned} \tag{30}$$

Accordingly, flow consistency in this paper refers to exact recomposition under the linear endpoint parameterization together with native-Euler recovery at zero correction. Local agreement with the learned vector field is reported as a separate empirical diagnostic in Appendix D.2.

G.3 Support-Aware Denominator Calibration

For a fixed packed identity gradient $\mathbf{s} \in \mathbb{R}^{N\times C}$ and soft support weights $\mathbf{w} \in [0,1]^N$, define

$$p = \frac{1}{N}\sum_i w_i, \qquad q = \frac{\sum_{i,c} w_i s_{i,c}^2}{\sum_{i,c} s_{i,c}^2}. \tag{31}$$

Using the whole-latent RMS and the support RMS in Eq. 12,

$$\mathrm{RMS}(\mathbf{s})^2 = \frac{\sum_{i,c} s_{i,c}^2}{CN}, \tag{32}$$

$$\mathrm{SRMS}(\mathbf{s};\mathbf{w})^2 = \frac{\sum_{i,c} w_i s_{i,c}^2}{C\sum_i w_i} = \frac{q}{p}\,\mathrm{RMS}(\mathbf{s})^2. \tag{33}$$

Therefore,

$$\frac{\mathrm{SRMS}(\mathbf{s};\mathbf{w})}{\mathrm{RMS}(\mathbf{s})} = \sqrt{\frac{q}{p}}. \tag{34}$$

For a prescribed support-local target magnitude T, whole-latent calibration gives

$$\gamma_{\mathrm{all}} = \frac{T}{\mathrm{RMS}(\mathbf{s})}, \qquad \gamma_{\mathrm{all}}\,\mathrm{SRMS}(\mathbf{s};\mathbf{w}) = T\sqrt{\frac{q}{p}}, \tag{35}$$

whereas support-aware calibration gives

$$\gamma_{\mathrm{sup}} = \frac{T}{\mathrm{SRMS}(\mathbf{s};\mathbf{w})}, \qquad \gamma_{\mathrm{sup}}\,\mathrm{SRMS}(\mathbf{s};\mathbf{w}) = T. \tag{36}$$

When the selected support contains nearly all identity-gradient energy, $q \simeq 1$, Eq. 34 reduces approximately to the familiar $1/\sqrt{p}$ scaling. For an unchanged gradient and unchanged support, appending zero-gradient background changes the whole-latent denominator while leaving the support-conditioned denominator unchanged. This identity characterizes the normalization step at a fixed target magnitude; the complete refinement pipeline additionally includes state evolution, decoding, acceptance, and spatial preservation.

H Additional Stability Properties

H.1 Candidate-Aligned Warm-Start Perturbation Bound

The spatial initialization in Eq. 2 does not claim that every spatial location corresponds to an independent flow time. Instead, all subsequent backbone evaluations use the shared scalar schedule τ_k. The spatial profile is only an initialization mechanism.

Consider a uniform reference initialization

$$\mathbf{h}_0^{\mathrm{uni}} = (1-\bar\rho)\mathbf{c} + \bar\rho\boldsymbol{\omega}, \tag{37}$$

for some scalar $\bar\rho$. The spatial initialization satisfies

$$\mathbf{h}_0(\mathbf{u}) - \mathbf{h}_0^{\mathrm{uni}}(\mathbf{u}) = [\rho(\mathbf{u}) - \bar\rho][\boldsymbol{\omega}(\mathbf{u}) - \mathbf{c}(\mathbf{u})]. \tag{38}$$

Hence

$$\left\|\mathbf{h}_0 - \mathbf{h}_0^{\mathrm{uni}}\right\| \le \|\rho - \bar\rho\|_\infty \left\|\boldsymbol{\omega} - \mathbf{c}\right\|. \tag{39}$$

Assume the frozen backbone displacement field is locally K-Lipschitz:

$$\|\mathcal{B}_\vartheta(\mathbf{y},\tau) - \mathcal{B}_\vartheta(\mathbf{y}',\tau)\| \le \mathsf{K}\|\mathbf{y} - \mathbf{y}'\|. \tag{40}$$

For the unguided update

$$\mathcal{T}_k(\mathbf{y}) = \mathbf{y} + (\tau_{k+1} - \tau_k)\mathcal{B}_\vartheta(\mathbf{y},\tau_k), \tag{41}$$

we have

$$\|\mathcal{T}_k(\mathbf{y}) - \mathcal{T}_k(\mathbf{y}')\| \le (1 + \mathsf{K}|\tau_{k+1} - \tau_k|)\,\|\mathbf{y} - \mathbf{y}'\|. \tag{42}$$

After K steps,

$$\|\mathbf{h}_K - \mathbf{h}_K^{\mathrm{uni}}\| \le \prod_{k=0}^{K-1} (1 + \mathsf{K}|\tau_{k+1} - \tau_k|)\,\|\mathbf{h}_0 - \mathbf{h}_0^{\mathrm{uni}}\|. \tag{43}$$

This establishes a finite-horizon perturbation bound for the spatial warm start under standard local regularity assumptions.

H.2 Accepted-Step Safety

Recall

$$\mathcal{H}_k(\mathbf{a}) = u_k \mathcal{J}_{\mathrm{ctx}}(\mathbf{a}) + s_k \mathcal{J}_{\mathrm{id}}(\mathbf{a}). \tag{44}$$

By construction, the backtracking rule accepts a candidate only if

$$\mathcal{H}_k(\bar{\mathbf{a}}_k) \le \mathcal{H}_k(\mathbf{a}_k) + \delta_{\mathrm{num}}, \tag{45}$$

$$\mathcal{J}_{\mathrm{id}}(\bar{\mathbf{a}}_k) \le \mathcal{J}_{\mathrm{id}}(\mathbf{a}_k) + \delta_{\mathrm{num}}. \tag{46}$$

If no candidate satisfies the criterion,

$$\bar{\mathbf{a}}_k = \mathbf{a}_k, \tag{47}$$

and both inequalities hold with equality. Thus, every accepted endpoint correction is non-increasing with respect to both the combined acceptance objective and the identity objective, up to numerical tolerance. This is a per-step clean-endpoint guarantee; it does not claim monotonic identity improvement across the entire subsequent generative trajectory.

H.3 Candidate-Region Fidelity Control

Define a trusted candidate region at threshold $\xi > 0$:

$$\mathcal{T}_\xi = \{\mathbf{u} : \varpi(\mathbf{u}) \ge \xi\}. \tag{48}$$

From Eq. 7,

$$\mathcal{J}_{\mathrm{ctx}} \ge \frac{\xi}{2} \sum_{\mathbf{u} \in \mathcal{T}_\xi} \|\mathbf{Y}(\mathbf{u}) - \mathbf{C}(\mathbf{u})\|_2^2. \tag{49}$$

Therefore,

$$\frac{1}{|\mathcal{T}_\xi|} \sum_{\mathbf{u} \in \mathcal{T}_\xi} \|\mathbf{Y}(\mathbf{u}) - \mathbf{C}(\mathbf{u})\|_2^2 \le \frac{2\mathcal{J}_{\mathrm{ctx}}}{\xi |\mathcal{T}_\xi|}. \tag{50}$$

Because $\mathcal{J}_{\mathrm{id}} \ge 0$ and $u_k > 0$,

$$\mathcal{J}_{\mathrm{ctx}}(\bar{\mathbf{a}}_k) \le \frac{\mathcal{H}_k(\bar{\mathbf{a}}_k)}{u_k}. \tag{51}$$

Combining this with Eq. 45,

$$\frac{1}{|\mathcal{T}_\xi|} \sum_{\mathbf{u} \in \mathcal{T}_\xi} \|\mathcal{V}(\bar{\mathbf{a}}_k)(\mathbf{u}) - \mathbf{C}(\mathbf{u})\|_2^2 \le \frac{2[\mathcal{H}_k(\mathbf{a}_k) + \delta_{\mathrm{num}}]}{u_k \xi |\mathcal{T}_\xi|}. \tag{52}$$

Equation 52 gives an explicit bound on candidate deviation in regions assigned sufficiently high retention weight. It should be interpreted as a pixel-space fidelity bound, not as a formal guarantee of semantic invariance.

H.4 Contraction of the Background Trajectory Pull

For a fixed location $\mathbf{u}$, Eq. 15 is

$$\mathbf{h}_{k+1} = (1 - \pi)\widetilde{\mathbf{h}}_{k+1} + \pi \mathbf{q}_{k+1}. \tag{53}$$

Subtracting the candidate trajectory gives

$$\mathbf{h}_{k+1} - \mathbf{q}_{k+1} = (1 - \pi)(\widetilde{\mathbf{h}}_{k+1} - \mathbf{q}_{k+1}). \tag{54}$$

Hence

$$\|\mathbf{h}_{k+1} - \mathbf{q}_{k+1}\| = (1 - \pi)\|\widetilde{\mathbf{h}}_{k+1} - \mathbf{q}_{k+1}\|. \tag{55}$$

For background locations with $\pi > 0$, the operation is therefore a strict contraction toward the candidate trajectory. For core and halo locations, $\pi = 0$, and Eq. 15 becomes the identity map:

$$\mathbf{h}_{k+1} = \widetilde{\mathbf{h}}_{k+1}. \tag{56}$$

This formalizes the different roles of the three spatial regions: candidate contraction is applied in the background while edit freedom is preserved within the protected facial neighborhood.

I Backbone Compatibility Interface

The main method is independent of the internal attention topology, but not independent of the underlying latent generative process. We therefore define a minimal compatibility interface. A refinement backbone must expose

$$\boxed{\text{state} \to \text{model displacement} \to \text{recoverable clean endpoint}} \tag{57}$$

together with a differentiable decoder and a consistent recomposition rule.

I.1 Joint-Attention Flow Backbone

For a FLUX-style instantiation, image conditions and the current state participate in the backbone's joint-attention computation. The implementation-specific adapter returns $\mathbf{d}_k$, after which the candidate refinement procedure in Sec. 3.2 is applied unchanged.

I.2 Causal-Reference Flow Backbone

For a Qwen-style instantiation, reference images populate causal reference states before target denoising. The target state subsequently reads these frozen reference features while predicting $\mathbf{d}_k$. Again, once the backbone prediction is obtained, endpoint recovery, decoder-mediated objectives, SAGC, backtracking, recomposition, and regional pull are unchanged.

Thus, the architectural difference is confined to the map

$$(\mathbf{h}_k, \tau_k, \boldsymbol{\psi}) \mapsto \mathbf{d}_k, \tag{58}$$

rather than to the identity-refinement mechanism itself.

J Implementation Details Corresponding to the Formulation

The current implementation uses a short decreasing flow schedule and keeps all generative and recognition networks frozen. Candidate initialization employs a larger perturbation in the face core and a near-clean perturbation elsewhere.

The face core and the wider halo are derived from the candidate-aligned target-face localization. The core controls the local initialization budget, while the wider halo controls where candidate-trajectory pull is disabled. These masks therefore have deliberately different roles.

The identity reference descriptor is computed once from an aligned reference face. The candidate alignment is also computed once and reused during clean-endpoint refinement. This avoids allowing changing landmark estimates to alter the identity objective itself.

At every refinement step, the identity and context gradients are obtained by differentiating through the frozen decoder. The identity gradient is calibrated using its face-support RMS, followed by a global RMS safety cap. A short backtracking sequence verifies candidate updates explicitly by decoding them and recomputing both objectives.

After accepted endpoint correction, the complementary endpoint recovered before optimization is retained exactly for recomposition. Candidate-trajectory pull is applied only after recomposition and is disabled throughout the core and halo. Its reference trajectory approaches the clean candidate as the terminal flow coordinate approaches zero.