FACEROUTE: SELECTIVE REFERENCE REWEIGHTING FOR MULTI-PERSON IDENTITY PRESERVATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-person image editing must preserve several facial identities while using full-image reference context. Frozen diffusion-transformer editors already encode the reference images, yet their generated faces can lose distinctive features or take on attributes of another person. We ask (i) whether stronger access to reference evidence improves identity preservation and (ii) whether this strengthening should be spatially selective. We propose FACEROUTE, a family of training-free routing variants that add a $\log \alpha$ bias from target-image queries to selected reference VAE keys, without changing backbone weights or adding a conditioning branch. On the two-person and three-/four-person splits of MultiID-Bench, whole-image routing (FACEROUTE-W) at $\alpha = 1.5$ improves reference and ground-truth identity similarity for both QWEN-IMAGE-EDIT-2511 and FLUX.2-KLEIN-BASE-9B, with all eight paired bootstrap intervals above zero. Face-local routing (FACEROUTE-F) also improves over both base editors in paired tests, and its two-person gains replicate across three seeds. Both variants also raise copy-paste preference and slightly lower aggregate aesthetic scores relative to the base editors; for whole-image routing, both costs grow when α is raised from 1.5 to 2.5. The answer to (ii) depends on reference face scale: on ordinary references, whole-image routing has higher paired identity similarity, and a mass-matched control links this advantage to the larger injected reference attention mass; on a small-face stress test, face-local routing scores higher on both identity metrics, with higher aesthetic scores and lower copy-paste preference.

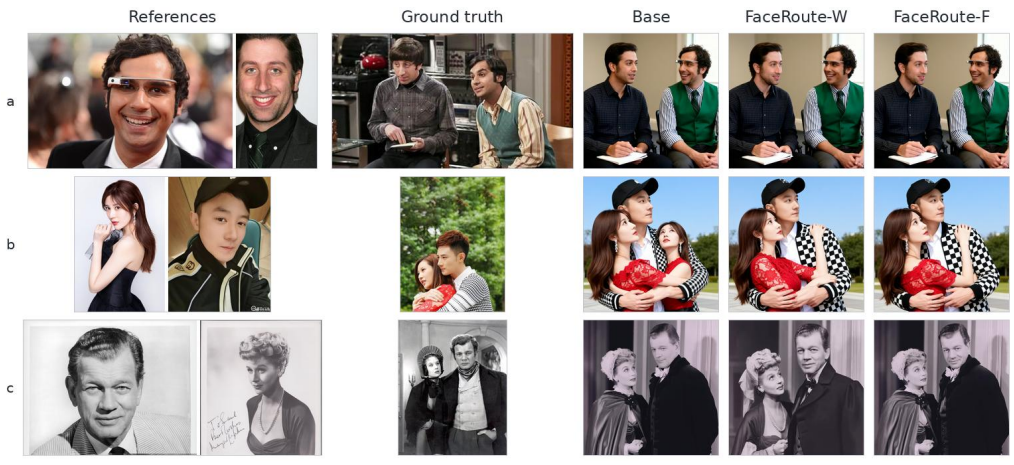

Figure 1: Three two-person MultiID-Bench examples with matched seeds and the 40-step QWEN-IMAGE-EDIT-2511 setting. Each row shows two identity references, ground truth, baseline, FACEROUTE-W and FACEROUTE-F; both routing methods use $\alpha = 1.5$.

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.

1 INTRODUCTION

Multi-person image editing asks a model to preserve several facial identities while composing a new scene. Frozen editors built on diffusion-transformer backbones (Peebles & Xie, 2023) encode full reference images and expose them to target-image queries through joint attention (Wu et al., 2025; Black Forest Labs, 2026), yet generated faces can lose distinctive features or acquire attributes of another person (Xiao et al., 2025; Gu et al., 2023; Xu et al., 2026) (Figure 1). The identity evidence is already in the conditioning sequence. We therefore ask (i) whether strengthening target-to-reference attention improves identity preservation, and (ii) whether this strengthening should be spatially selective.

For (i), we add a constant $\log \alpha$ to the attention logits from every target-image query to the reference-image VAE keys. This multiplies their unnormalized attention weights by α, without training, weight changes, or a new conditioning branch. Applied to all reference keys, this whole-image routing (FACEROUTE-W) at $\alpha = 1.5$ improves reference and ground-truth identity similarity with frozen QWEN-IMAGE-EDIT-2511 (Wu et al., 2025; Qwen Team, 2025) and FLUX.2-KLEIN-BASE-9B (Black Forest Labs, 2026) on the two-person (P2) and three-/four-person (P3) splits of MultiID-Bench (Xu et al., 2026); all eight paired bootstrap intervals lie above zero. The ground-truth photograph is never given to the editor, so its similarity gain indicates recovered identity rather than only closer reproduction of the references. On Qwen P2, adding the same offset to text keys instead lowers both identity metrics, tying the gain to the reference pathway rather than to a generic positive bias. Under its default attention allocation, the editor thus under-uses reference evidence that helps identity. The gain has a cost: copy-paste preference rises, and raising α further trades aesthetic score for identity.

Whole-image routing, however, raises every reference key equally: face, body, and background. On MultiID-Bench P2, detected face boxes cover 16.35% of the reference area on average, so most boosted keys lie outside the face, and this imbalance grows for full-body or environmental references. To study this regime, we assemble a small-face stress test following the MultiID-Bench protocol (Xu et al., 2026), where the mean detected face area drops to 1.63%. Face-local routing (FACEROUTE-F) detects each reference face with SCRFD (Guo et al., 2022), maps the box to packed VAE tokens, and adds the same $\log \alpha$ only to those keys. All other reference tokens remain in the sequence, so the editor keeps appearance and scene context. The two variants share one reweighting rule and differ only in the selected key set.

The answer to (ii) depends on face scale. On ordinary P2 references, FACEROUTE-W has higher paired reference similarity than FACEROUTE-F on both backbones, and a mass-matched control links this advantage to injected attention mass: at equal injected reference mass, neither backbone shows a resolved identity difference. On the small-face stress test, the ordering reverses: FACEROUTE-F scores higher on both identity metrics with QWEN-IMAGE-EDIT-2511, and on a backbone-aligned subset, with both backbones, with higher aesthetic scores and lower copy-paste preference.

Our contributions are as follows.

- **Reference access as a training-free identity control.** One logit offset on target-to-reference attention improves reference and ground-truth identity similarity in two frozen editors on both MultiID-Bench splits; a text-key control ties this gain to the reference pathway. Face-local P2 gains replicate across three seeds, and α exposes an identity–quality trade-off.
- **A face-scale-dependent choice of routing scope.** Whole-image routing is stronger on ordinary references, an edge a mass-matched control links to injected attention mass; face-local routing is stronger on small-face references.

Replacing face boxes with SAM3 (Carion et al., 2026) subject masks transfers the key-selection rule to XVerseBench (Chen et al., 2025) (300 prompts, FLUX), where subject and identity similarity improve over the base editor.

2 RELATED WORK

Reference-conditioned personalization. Personalization methods use fine-tuning, learned embeddings, adapters, or identity encoders to represent a subject (Ruiz et al., 2023; Gal et al., 2023; Kumari

et al., 2023; Ye et al., 2023; Wang et al., 2024; Li et al., 2024; Guo et al., 2024). Multi-subject systems introduce grounding, regional attention, interaction modules, two-phase customization (Avrahami et al., 2023), or staged sampling (Wang et al., 2025; Kim et al., 2024; Kong et al., 2024; Gu et al., 2023; Xiao et al., 2025; Ma et al., 2024; She et al., 2026; Xu et al., 2026; Hu et al., 2025; Zhou et al., 2024). PAF combines inversion, token replacement, perturbation, and masks (Feng et al., 2026). FACEROUTE instead reweights reference VAE tokens already present in a frozen editor.

Inference-time attention control. Prompt-to-Prompt edits text–image cross-attention, MasaCtrl reuses source self-attention keys and values, and FreeControl guides diffusion features with semantic subspaces and structural conditions (Hertz et al., 2023; Cao et al., 2023; Mo et al., 2024); related work also injects source features (Tumanyan et al., 2023). These methods address source-image editing or structure-conditioned generation. FACEROUTE changes target-to-reference logits in multi-reference joint attention. Appendix Y compares these interfaces, including automatically derived masks where supported.

DeContext (Shen et al., 2025) optimizes pixel perturbations to suppress target-to-context attention and protect images from unauthorized editing. We intervene directly on the same attention pathway in the opposite direction: increasing access to selected reference keys while retaining unperturbed RGB inputs.

3 PROBLEM SETUP

Consider one denoising step of a frozen editor conditioned on a text prompt and N reference images, one per identity to be preserved. Let H^{text}, H^{tgt}, and H_i^{ref} denote the prompt-encoder tokens (multimodal for QWEN-IMAGE-EDIT-2511), the noisy target-image tokens, and the packed VAE tokens of reference i. Their joint-attention sequence of length L is

$$H = [H^{\text{text}};\ H^{\text{tgt}};\ H_1^{\text{ref}};\ \ldots;\ H_N^{\text{ref}}]. \tag{1}$$

We write $\mathcal{T} \subset \{1, \ldots, L\}$ for the positions of H^{tgt} and $\mathcal{R}_i$ for those of H_i^{ref}. For one head with key dimension d, let Q, K, V be the projections of H; the attention logits and output are

$$A = \frac{QK^\top}{\sqrt{d}} + M, \qquad O = \text{softmax}(A)V, \tag{2}$$

where M is the backbone's existing mask. Target queries $q \in \mathcal{T}$ already attend to every $\mathcal{R}_i$ in the baseline. We ask whether changing only the logits from $\mathcal{T}$ to a selected key set $\mathcal{K} \subseteq \bigcup_i \mathcal{R}_i$ improves identity preservation while following the prompt, with backbone weights fixed and without specifying where each person should appear.

Similarity to a reference can increase through better identity retention or through more literal reproduction of the reference photo. We therefore measure identity against both the references (sim_{ref}) and the benchmark ground truth (sim_{GT}), a different photo of each person that is not given to the editor, and use the copy-paste preference M_{CP} (Xu et al., 2026) (Eq. 6) to quantify relative preference for the reference over the ground truth.

4 METHOD

FACEROUTE strengthens target-to-reference attention with a single logit bias. FACEROUTE-W applies it to all reference keys, and FACEROUTE-F restricts it to detected face keys; the two variants differ only in the selected key set.

4.1 REWEIGHTING REFERENCE ACCESS

For a joint sequence of length L, let $\mathcal{K}$ be the set of selected reference keys. Define a head-shared bias $B^{(\ell,t)} \in \mathbb{R}^{L\times L}$ at layer ℓ and denoising step t:

$$B_{qk}^{(\ell,t)} = \log \alpha \, \mathbb{1}[q \in \mathcal{T},\ k \in \mathcal{K}], \tag{3}$$

$$O' = \text{softmax}(A + B)V. \tag{4}$$

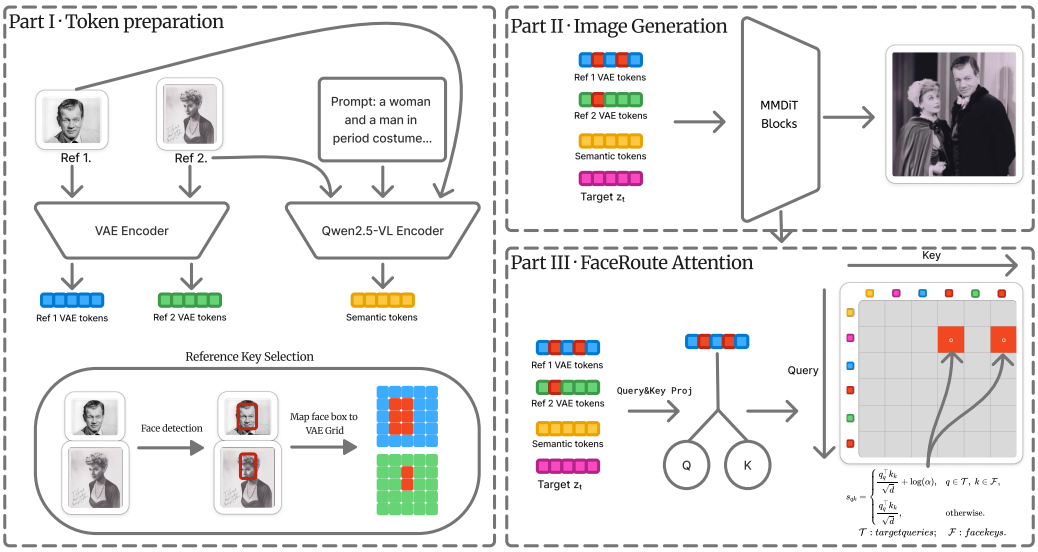

Figure 2: FACEROUTE-F in Qwen-Image-Edit-2511. Qwen2.5-VL provides multimodal conditioning, while the VAE supplies reference-image and target tokens. Detected face boxes select reference keys $\mathcal{F}$; the whole-image variant instead selects all reference keys (Section 4.2). The reused DiT attention operation adds $\log \alpha$ from target queries to the selected keys; no additional K/V branch or trainable weights are introduced.

We denote the all-target query scope $q \in \mathcal{T}$ by $\mathcal{T}_{\text{all}}$. Adding $\log \alpha$ multiplies the unnormalized attention weight of each selected key by α; $\alpha = 1$ recovers the baseline. The default uses a constant α across all layers and steps, introducing one user-controlled hyperparameter but no trainable weights.

For one target query, let $S = \sum_{k \in \mathcal{K}} \text{softmax}(A)_{qk}$ be its original selected-key mass. At fixed logits, the biased mass is exactly

$$S' = \frac{\alpha S}{1 + (\alpha - 1)S}. \tag{5}$$

When $\alpha > 1$ and $0 < S < 1$, the selected-key mass increases while relative weights within $\mathcal{K}$ are preserved.

4.2 WHOLE-IMAGE SELECTION

The simplest instantiation sets $\mathcal{K}$ to all packed VAE tokens of every reference image, $\mathcal{K} = \bigcup_{i=1}^{N} \mathcal{R}_i$. Every target query then reads all reference content more strongly. This variant requires no face detector and serves as both a standalone method and a comparator for spatial selection.

4.3 FACE-LOCAL SELECTION

FACEROUTE-F narrows $\mathcal{K}$ to tokens that correspond to detected faces (Figure 2). For each reference I_i, SCRFD (Guo et al., 2022) detects its face on the image resized to the VAE canvas resolution. Detections covering less than 0.5% of the canvas are discarded. The retained bounding box is mapped to packed VAE token indices using floor and ceiling boundaries, producing an index set $\mathcal{F}_i \subseteq \mathcal{R}_i$. If no eligible face is found, $\mathcal{F}_i$ is empty and that reference receives no bias. The union $\mathcal{F} = \bigcup_{i=1}^{N} \mathcal{F}_i$ gives the face-local key set $\mathcal{K} = \mathcal{F}$.

Every target query receives the same bias toward $\mathcal{K}$. The mask is symmetric over reference order, but learned logits still determine which reference a target person prefers. Layer schedules and cross-reference isolation are diagnostic variants tested in Appendix M.

5 EXPERIMENTAL SETUP

5.1 BACKBONES AND DATA

We evaluate QWEN-IMAGE-EDIT-2511 (Wu et al., 2025; Qwen Team, 2025) (hereafter Qwen) and FLUX.2-KLEIN-BASE-9B (Black Forest Labs, 2026) (hereafter FLUX) with 40 denoising steps, guidance scale 4, and 1024×1024 outputs. Controlled comparisons are within-backbone: each routed output shares the prompt, references, and seed of its baseline. Unless stated otherwise, routing uses $\alpha = 1.5$ with all-target queries ($\mathcal{T}_{\text{all}}$).

MultiID-Bench (Xu et al., 2026) provides 194 two-person (P2) and 33 three-/four-person (P3) instances; a few P3 analyses use an earlier 31-ID cohort, marked where they appear (Appendix B). As system-level context, we report the training-free PAF (Feng et al., 2026) and three trained methods: XVerse (Chen et al., 2025), MOSAIC (She et al., 2026), and WithAnyone (Xu et al., 2026). Their backbones and protocols differ, so these rows are not controlled comparisons (Appendix Y).

For the scope ablation, we collect a 49-example two-person small-face stress test from the web with the MultiID-Bench collection procedure (Xu et al., 2026). Examples are not selected by routing outcome; the set size is limited by data availability. Detected faces cover 1.63% of the reference area on average, versus 16.35% on ordinary P2 references. All 49 examples are evaluated with Qwen; an aligned 31-example subset is also evaluated with FLUX (Appendix M.4). Implementation details are in Appendix F.

5.2 METRICS AND STATISTICAL PROTOCOL

Following the released MultiID-Bench evaluator, generated faces are matched to ground-truth and reference faces by two independent Hungarian assignments; a reference-anchored audit on Qwen preserves the direction of the identity and CP changes (Appendix D). The primary scores, $\mathrm{sim}_{\mathrm{ref}}$ and $\mathrm{sim}_{\mathrm{GT}}$, are cosine similarities to the matched reference and ground-truth faces under ArcFace, FaceNet, and AdaFace (Deng et al., 2019; Schroff et al., 2015; Kim et al., 2022). CLIP-I and CLIP-T measure image-level ground-truth similarity and prompt alignment (Radford et al., 2021). Bld (blending) is the mean identity similarity to non-matched ground-truth persons; lower is better. Aes is the aesthetic score (discus0434, 2024; Zhai et al., 2023). For a matched generated, ground-truth, and reference triplet (g, t, r), the copy-paste preference (CP) (Xu et al., 2026) is

$$\mathrm{M}_{\mathrm{CP}}(g,t,r) = \begin{cases} 0, & \theta_\varepsilon(t,r) < \varepsilon, \\ \dfrac{\theta_\varepsilon(g,t) - \theta_\varepsilon(g,r)}{\theta_\varepsilon(t,r)}, & \text{otherwise}, \end{cases} \tag{6}$$

where $\varepsilon = 10^{-6}$ and θ_ε is the angular distance computed with stabilized embedding normalization and cosine clipping (Appendix F.1); larger values indicate greater reference preference.

Paired comparisons average encoder scores within each example, without outlier filtering, and report percentile 95% bootstrap intervals (10,000 resamples over exact common IDs); they are the basis for our MultiID-Bench conclusions. Table 1 and the 49-example comparison instead pool encoder scores under the evaluator's $1.5\times$IQR filter, without intervals, and are descriptive; the 31-example tables use unfiltered means (Appendices B and G).

The default $\alpha = 1.5$ was chosen retrospectively from a Qwen sweep on the same benchmark (Table 18) and reused unchanged for both routing variants and FLUX. Split-half checks recover it in 93.9% of P2 folds and 51.5% of P3 folds (Appendix M.2). FID (Heusel et al., 2017) is reported only in appendix tables.

6 MAIN RESULTS

6.1 WHOLE-IMAGE ROUTING

At $\alpha = 1.5$, FACEROUTE-W improves identity on every backbone–split combination: paired $\Delta\mathrm{sim}_{\mathrm{ref}}$ ranges from $+0.047$ to $+0.074$ and $\Delta\mathrm{sim}_{\mathrm{GT}}$ from $+0.024$ to $+0.034$, with all eight 95% bootstrap intervals above zero (Table 46).

Table 1: MultiID-Bench results. FACEROUTE-W biases all reference VAE keys; FACEROUTE-F biases detected-face keys only; the number after each routed system is α. Scores are method-wise pooled, IQR-filtered aggregates of available scores in each split. Cross-system backbones and protocols differ (Appendix Y). Dark/light yellow and green mark the best/second-best trained and training-free systems within each split and metric. PAF† requires additional layout masks; ‡its CP uses 68–70 P2 and 5 P3 samples per encoder before IQR filtering.

System	Processed / Total	sim_{GT} ↑	sim_{ref} ↑	CP ↓	Bld ↓	CLIP-I ↑	CLIP-T ↑	Aes ↑
Two-person split — Trained personalization								
XVerse	194/194	0.2298	0.2742	0.0374	0.0870	0.7276	0.3218	5.4623
MOSAIC	194/194	0.2356	0.3113	0.0723	0.0929	0.6970	0.3288	5.7693
WithAnyone	194/194	0.4423	0.6397	0.2235	0.0688	0.7638	0.3203	4.9479
Two-person split — Training-free inference								
PAF†	194/194	0.2486	0.3962	0.1465‡	0.0580	0.6639	0.2961	4.5769
Qwen Base	194/194	0.3892	0.5466	0.1848	0.1193	0.7766	0.3310	5.5301
Qwen + FACEROUTE-W 1.5	194/194	0.4214	0.6056	0.2289	0.1140	0.7804	0.3288	5.4815
Qwen + FACEROUTE-W 2.5	194/194	0.4384	0.6406	0.2549	0.1145	0.7806	0.3296	5.3937
Qwen + FACEROUTE-F 1.5	194/194	0.4170	0.5916	0.2129	0.1168	0.7768	0.3318	5.5003
Qwen + FACEROUTE-F 2.5	194/194	0.4362	0.6359	0.2473	0.1143	0.7800	0.3302	5.4834
FLUX.2 Base	194/194	0.3465	0.4388	0.1044	0.0775	0.7527	0.3365	5.8342
FLUX.2 + FACEROUTE-W 1.5	194/194	0.3766	0.4866	0.1236	0.0746	0.7611	0.3358	5.7901
FLUX.2 + FACEROUTE-W 2.5	194/194	0.3979	0.5376	0.1658	0.0756	0.7594	0.3309	5.7303
FLUX.2 + FACEROUTE-F 1.5	194/194	0.3684	0.4789	0.1204	0.0746	0.7553	0.3359	5.8098
FLUX.2 + FACEROUTE-F 2.5	194/194	0.3907	0.5271	0.1608	0.0809	0.7595	0.3343	5.8119
Three-/four-person split — Trained personalization								
XVerse	33/33	0.2196	0.2687	0.0449	0.0746	0.7157	0.3266	5.5635
MOSAIC	33/33	0.2353	0.3116	0.0693	0.0846	0.6736	0.3179	5.4441
WithAnyone	33/33	0.4605	0.6561	0.2413	0.0496	0.7647	0.3280	4.8973
Three-/four-person split — Training-free inference								
PAF†	31/33	0.2627	0.3788	0.1153‡	0.0565	0.6909	0.3141	4.5582
Qwen Base	33/33	0.3785	0.5314	0.1958	0.0843	0.7647	0.3212	5.5000
Qwen + FACEROUTE-W 1.5	33/33	0.4021	0.6051	0.2550	0.0795	0.7560	0.3207	5.4556
Qwen + FACEROUTE-W 2.5	33/33	0.4323	0.6681	0.2984	0.0767	0.7671	0.3150	5.2661
Qwen + FACEROUTE-F 1.5	33/33	0.4049	0.6057	0.2354	0.0810	0.7748	0.3180	5.4366
Qwen + FACEROUTE-F 2.5	33/33	0.4159	0.6543	0.2998	0.0778	0.7670	0.3151	5.3975
FLUX.2 Base	33/33	0.3044	0.3626	0.0759	0.0465	0.7128	0.3388	6.4927
FLUX.2 + FACEROUTE-W 1.5	33/33	0.3383	0.4247	0.0981	0.0445	0.7186	0.3404	6.4198
FLUX.2 + FACEROUTE-W 2.5	33/33	0.3919	0.5066	0.1400	0.0547	0.7353	0.3360	6.2450
FLUX.2 + FACEROUTE-F 1.5	33/33	0.3391	0.4215	0.1014	0.0436	0.7157	0.3402	6.4741
FLUX.2 + FACEROUTE-F 2.5	33/33	0.3730	0.4839	0.1463	0.0472	0.7300	0.3349	6.3854

For FACEROUTE-W, raising α from 1.5 to 2.5 further increases identity similarity (differences between the 2.5 and 1.5 aggregates in Table 1: $\Delta\text{sim}_{\text{ref}}$ $+0.0350$ to $+0.0819$; $\Delta\text{sim}_{\text{GT}}$ $+0.0170$ to $+0.0536$), while CP increases and Aes decreases on all four backbone–split combinations, showing an identity–quality trade-off between these two strengths.

6.2 FACE-LOCAL ROUTING

Restricting the bias to detected-face keys retains most of the identity gain of FACEROUTE-W. On Qwen P2, FACEROUTE-F at $\alpha = 1.5$ improves sim_{GT} over Base by a paired $+0.0262$ $[+0.0212, +0.0313]$ (Table 7) and has lower CP than FACEROUTE-W at the same strength (paired FACEROUTE-W$-$FACEROUTE-F $\Delta\text{M}_{\text{CP}} = +0.0115$ $[+0.0036, +0.0196]$); this paired reduction also holds on Qwen P3 (Table 3).

System context. Table 1 also includes XVerse, MOSAIC, WithAnyone, and PAF as system context; they are not controlled comparisons (Section 5). Qwen FACEROUTE-F at $\alpha = 2.5$ reaches P2

identity scores close to WithAnyone's, $(\mathrm{sim}_{\mathrm{ref}}, \mathrm{sim}_{\mathrm{GT}}) = (0.6359, 0.4362)$ versus $(0.6397, 0.4423)$, without training, while WithAnyone has lower CP (0.2235 vs. 0.2473).

6.3 Transfer to Subject Masks

Replacing face boxes with SAM3 (Carion et al., 2026) subject masks transfers the same key-selection rule beyond faces. On all 300 XVerseBench prompts (Chen et al., 2025), covering human, animal, and object subjects, on FLUX with the settings of Section 5 and seed 42, $\alpha = 1.5$ raises non-human subject similarity (IP) from 66.59 to 68.11 and human identity (ID) from 50.12 to 54.68, maintains DPG ($94.42 \rightarrow 94.93$), and slightly lowers AES ($62.92 \rightarrow 62.44$). At $\alpha = 2.5$, IP and ID rise further (69.65, 60.09) while DPG and AES decrease, following the identity–quality trade-off observed on MultiID-Bench. XVerse and MOSAIC scores, per-subject-count breakdowns, and decay schedules appear in Appendix P.

7 Ablation Studies

Strength sensitivity is reported in Section 6.1. Here we ablate reference scope, seed stability, and pathway controls.

7.1 Effect of Routing Scope

We test whether restricting the bias to face keys changes the outcome. Paired comparisons use 194 P2 and 33 P3 examples.

Ordinary references. On Qwen P2, detected face boxes cover 16.35% of the reference canvas on average. At this scale, FaceRoute-W holds a small edge: the paired $\Delta\mathrm{sim}_{\mathrm{ref}}$ is $+0.011$ for Qwen and $+0.008$ for FLUX, with $\Delta\mathrm{sim}_{\mathrm{GT}}$ differences of $+0.004$ and $+0.007$ respectively (Table 2). When faces occupy a substantial fraction of the image, whole-image routing applies the bias to more reference keys while still covering the face region.

Small reference faces. The identity ordering reverses on the small-face stress set. On this 49-example set where face boxes occupy only 1.63% of the reference canvas, FaceRoute-F has higher scores than FaceRoute-W on both identity metrics: $+0.0086$ in $\mathrm{sim}_{\mathrm{ref}}$ and $+0.0024$ in $\mathrm{sim}_{\mathrm{GT}}$ on Qwen (Figure 3; Appendix M.4).

The aligned 31-example cross-backbone subset shows the same FaceRoute-F-over-FaceRoute-W ordering: $+0.0083/+0.0035$ in $\mathrm{sim}_{\mathrm{ref}}/\mathrm{sim}_{\mathrm{GT}}$ for Qwen and $+0.0046/+0.0075$ for FLUX. CP is lower by 0.0072 and 0.0061, and the aesthetic score is higher by 0.041 and 0.085, respectively (Tables 19–20).

Reference-background transfer. We randomly sampled 80 eligible P2 pairs and included 31 P3 pairs for a blinded, single-rater comparison of prompt-unrequested reference-background transfer at $\alpha = 1.5$. Among the 16 of 111 pairs for which the rater identified a direction, FaceRoute-W showed more transfer in 14 and FaceRoute-F in 2. Figure 11 shows three selected P2 cases; the rating protocol and full counts are in Appendix M.5.

7.2 Multi-Seed Replication

At $\alpha = 1.5$, FaceRoute-F–Base paired effects on P2 replicate across three generation seeds on both backbones (Tables 8 and 11): $\Delta\mathrm{sim}_{\mathrm{ref}}$ ranges from $+0.039$ to $+0.045$ and $\Delta\mathrm{sim}_{\mathrm{GT}}$ from $+0.020$ to $+0.026$, with every per-seed interval above zero. Per-seed CP results, including CP-specific encoders and valid counts, appear in Appendix G.

7.3 Pathway and Attention-Mass Controls

Applying the same $\log 1.5$ offset to text keys instead of reference keys *decreases* identity similarity on Qwen P2 ($\Delta\mathrm{sim}_{\mathrm{ref}} = -0.048$, $\Delta\mathrm{sim}_{\mathrm{GT}} = -0.022$), whereas face-key routing raises both ($+0.045$

Table 2: FACEROUTE-W minus FACEROUTE-F paired effects at $\alpha = 1.5$. Positive values favor whole-image scope. Intervals are 95% percentile-bootstrap CIs (10,000 resamples) on untrimmed exact-ID paired differences ($n = 194$ for P2 and $n = 33$ for P3; Appendix C).

Backbone / split	$\Delta\text{sim}_{\text{ref}} \uparrow$	$\Delta\text{sim}_{\text{GT}} \uparrow$
Qwen / P2	$+0.0113$ [+0.0031, +0.0202]	$+0.0043$ [-0.00004, +0.0087]
Qwen / P3	$+0.0147$ [-0.0022, +0.0306]	-0.0027 [-0.0112, +0.0050]
FLUX / P2	$+0.0077$ [+0.0029, +0.0126]	$+0.0074$ [+0.0035, +0.0114]
FLUX / P3	$+0.0033$ [-0.0098, +0.0161]	-0.0009 [-0.0100, +0.0076]

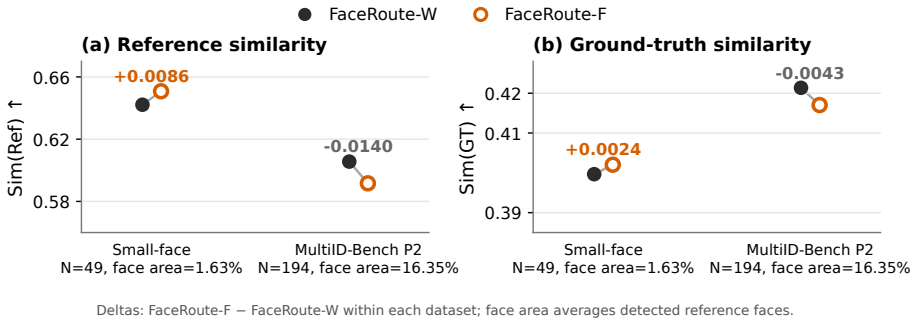

Figure 3: Scope comparison on Qwen at $\alpha = 1.5$. FACEROUTE-F has higher identity scores on the 49-example small-face stress test; FACEROUTE-W is higher on ordinary P2. Labels give sample counts and mean detected face area. Both datasets use method-wise pooled encoder scores with metric-wise IQR filtering. Deltas are FACEROUTE-F minus FACEROUTE-W within each dataset (Appendix M.4).

and $+0.026$). The gain therefore depends on the reference pathway rather than on a positive logit offset alone (Appendix W).

A mass-matched control on ordinary P2 adjusts whole-image bias strength to match the reference attention mass of face-local routing at fixed logits, per query and head. The mass-matched–minus–face-local differences are unresolved on both backbones ($\Delta\text{sim}_{\text{ref}}$: -0.0025 for Qwen, -0.0027 for FLUX; $\Delta\text{sim}_{\text{GT}}$: -0.0014 and -0.0010; all intervals span zero), while whole-image routing at fixed α retains a significant identity advantage over the mass-matched arm (Appendix S). This links the equal-α scope difference on ordinary P2 references to injected reference mass. The small-face advantage of FACEROUTE-F is consistent with concentrating the bias on identity-bearing tokens (Section 7.1). Additional controls on detector-box margins, reference isolation, and layer schedules appear in Appendix M; identity-separation, completeness, position-binding, and copy audits appear in Appendices O, R, H, N, and U.

8 DISCUSSION

Under default attention allocation, both tested editors under-use reference evidence that helps identity: a training-free logit bias toward reference VAE keys raises identity similarity on both backbones, whereas the same offset on text keys lowers it on Qwen P2. Whether the bias should be spatially selective depends on face scale. On ordinary P2 references, no resolved identity difference remains once FACEROUTE-W is rescaled to inject the same reference attention mass as FACEROUTE-F; on the small-face stress set, FACEROUTE-F, which concentrates the bias on face keys, scores higher on both identity metrics (Section 7.1).

The intervention changes only attention logits between existing tokens. It needs no training, weight changes, or target-side layout masks; FACEROUTE-F additionally needs only reference face boxes. It therefore applies to editors that expose compatible target-to-reference joint-attention logits. Routing strength sets an identity–quality trade-off: for FACEROUTE-W, raising α from 1.5 to 2.5 increases identity similarity and CP and lowers Aes on both backbones and splits (Table 1).

Future directions. The fixed scalar could be replaced by a lightweight learned bias module trained with the backbone frozen. As a first step beyond image editing, a two-pass, trajectory-guided adaptation to the MiniMax-H3 image-to-video model (MiniMax, 2026) raises video-mean ArcFace similarity by $+0.0268$ and last-frame similarity by $+0.0434$ on ten single-person inputs (Appendix X).

9 LIMITATIONS

FACEROUTE operates on keys the backbone has already encoded; it cannot recover identity detail absent from the reference representation or guarantee correct identity-to-position binding when the backbone confuses target persons. Increasing α can improve identity similarity at an aesthetic cost (Table 1), and the current method offers no automatic way to select the optimum.

The face-local identity advantage is observed on the small-face stress set. On ordinary P2, whole-image routing has higher paired $\mathrm{sim}_{\mathrm{ref}}$ on both backbones, while P3 identity-scope intervals span zero (Table 2). A face-area threshold for choosing between the scopes remains to be determined.

Our identity evaluation relies on embedding-based metrics. Image experiments cover two backbones; the video extension evaluates ten single-person H3 inputs under a face-visible, two-pass protocol (Appendix X). Generalization to other architectures and broader motion conditions remains to be evaluated.

10 CONCLUSION

Frozen multi-person editors can make better use of the reference evidence they already encode. With backbone, prompts, reference images, and seeds fixed, FACEROUTE-W improves identity similarity over Base on both tested editors, at the cost of higher CP and slightly lower Aes. The logit bias requires no training, architectural changes, or target-side layout masks. The best key scope depends on face scale: on ordinary references, the advantage of FACEROUTE-W tracks injected reference attention mass, whereas on the 49-example small-face stress set, FACEROUTE-F has higher identity similarity than FACEROUTE-W.

AI USE STATEMENT

Large language models assisted with literature discovery, drafting and revising text, implementing and debugging code, experiment planning, and interpreting results and reviewer feedback. The core research idea originated with the authors, who retain responsibility for the implementation, source verification, experimental evidence, and all claims. AI assistance is not an independent validation of those claims.

ETHICS STATEMENT

Identity-preserving editing can support creative and assistive work but also facilitate impersonation, non-consensual editing, and deceptive media. Deployments should require rights or consent for references, preserve provenance where possible, and restrict high-risk identity use. Face encoders may carry demographic bias, so aggregate similarity does not imply uniform performance. MultiID-Bench is used under its non-commercial research license; we do not redistribute source images. Public availability does not establish subject consent, and the benchmark lacks demographic labels for responsible subgroup analysis. The video extension carries the same identity-misuse risks.

REPRODUCIBILITY STATEMENT

The method, inference settings, metric definitions, valid-set rules, and paired analyses are specified in the main text and appendices. The appendix includes per-seed results and the two-pass H3 protocol. Code, evaluation scripts, and experiment configurations are planned for release upon acceptance; Appendix F distinguishes recorded details from remaining immutable-artifact gaps. No public code release is claimed here.

REFERENCES

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-A-Scene: Extracting multiple concepts from a single image. In *ACM SIGGRAPH Asia 2023 Conference Papers*, pp. 1–12, 2023. doi: 10.1145/3610548.3618154.

Black Forest Labs. FLUX.1-dev model card. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024. Official model card; not peer-reviewed.

Black Forest Labs. FLUX.2-klein-base-9B model card. https://huggingface.co/black-forest-labs/FLUX.2-klein-base-9B, 2026. Official model card; not peer-reviewed.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22560–22570, 2023. doi: 10.1109/ICCV51070.2023.02062.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e0982cbc81401df3430ee1ff780dc7a2-Abstract-Conference.html.

Bowen Chen, Brynn Zhao, Haomiao Sun, Li Chen, Xu Wang, Daniel Du, and Xinglong Wu. XVerse: Consistent multi-subject control of identity and semantic attributes via DiT modulation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 7437–7459, 2025. doi: 10.52202/085713-0254. URL https://papers.nips.cc/paper_files/paper/2025/hash/0b0153a91f827b14e8bfea4e211362f3-Abstract-Conference.html.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019. doi: 10.1109/CVPR.2019.00482.

discus0434. Aesthetic predictor v2.5. https://github.com/discus0434/aesthetic-predictor-v2-5, 2024. SigLIP-based aesthetic scoring model.

Haoran Feng, Zehuan Huang, Lin Li, and Lu Sheng. Personalize anything for free with diffusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3921–3929, 2026. doi: 10.1609/aaai.v40i5.37394.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NAQvF08TcyG.

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Ying Shan, and Mike Zheng Shou. Mix-of-Show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15890–15902, 2023. doi: 10.52202/075280-0699.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RhB1AdoFfGE.

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, pp. 36777–36804, 2024. doi: 10.52202/079017-1159.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_CDixzkzeyb.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

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.

Chanran Kim, Jeongin Lee, Shichang Joung, Bongmo Kim, and Yeul-Min Baek. InstantFamily: Masked attention for zero-shot multi-id image generation. arXiv preprint arXiv:2404.19427, 2024. URL https://arxiv.org/abs/2404.19427.

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*, pp. 18750–18759, 2022. doi: 10.1109/CVPR52688.2022.01819.

Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. OMG: Occlusion-friendly personalized multi-concept generation in diffusion models. In *European Conference on Computer Vision*, pp. 253–270, 2024. doi: 10.1007/978-3-031-72751-1_15.

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.

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.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-Diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024. doi: 10.1145/3641519.3657469.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025. doi: 10.1109/ICCV51701.2025.01400. URL https://openaccess.thecvf.com/content/ICCV2025/html/Ma_HPSv3_Towards_Wide-Spectrum_Human_Preference_Score_ICCV_2025_paper.html.

MiniMax. MiniMax-H3 model card. https://huggingface.co/MiniMaxAI/MiniMax-H3, 2026. Official model card; not peer-reviewed.

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.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. doi: 10.1109/ICCV51070.2023.00387.

Qwen Team. Qwen-Image-Edit-2511 model card. https://huggingface.co/Qwen/Qwen-Image-Edit-2511, 2025. Official model card; not peer-reviewed.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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.

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*, pp. 815–823, 2015. doi: 10.1109/CVPR.2015.7298682.

Dong She, Siming Fu, Mushui Liu, Qiaoqiao Jin, Hualiang Wang, Weihao Zhou, and Jidong Jiang. MOSAIC: Multi-subject personalized generation via correspondence-aware alignment and disentanglement. In *International Conference on Learning Representations*, pp. 63128–63156, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/67182d38f943d51ff5f1896841751243-Abstract-Conference.html.

Linghui Shen, Mingyue Cui, and Xingyi Yang. DeContext as defense: Safe image editing in diffusion transformers. arXiv preprint arXiv:2512.16625, 2025. URL https://arxiv.org/abs/2512.16625.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1921–1930, 2023. doi: 10.1109/CVPR52729.2023.00191.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. InstantID: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519, 2024. URL https://arxiv.org/abs/2401.07519.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. MS-Diffusion: Multi-subject zero-shot image personalization with layout guidance. In *International Conference on Learning Representations*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. arXiv preprint arXiv:2508.02324, 2025. URL https://arxiv.org/abs/2508.02324.

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(3):1175–1194, 2025. doi: 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: Toward controllable and id consistent image generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d432fbe4877ee1a6a51632a18e69784f-Abstract-Conference.html.

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

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023. doi: 10.1109/ICCV51070.2023.01100.

Zhengguang Zhou, Jing Li, Huaxia Li, Nemo Chen, and Xu Tang. StoryMaker: Towards holistic consistent characters in text-to-image generation. arXiv preprint arXiv:2409.12576, 2024. URL https://arxiv.org/abs/2409.12576.

A Dataset Examples

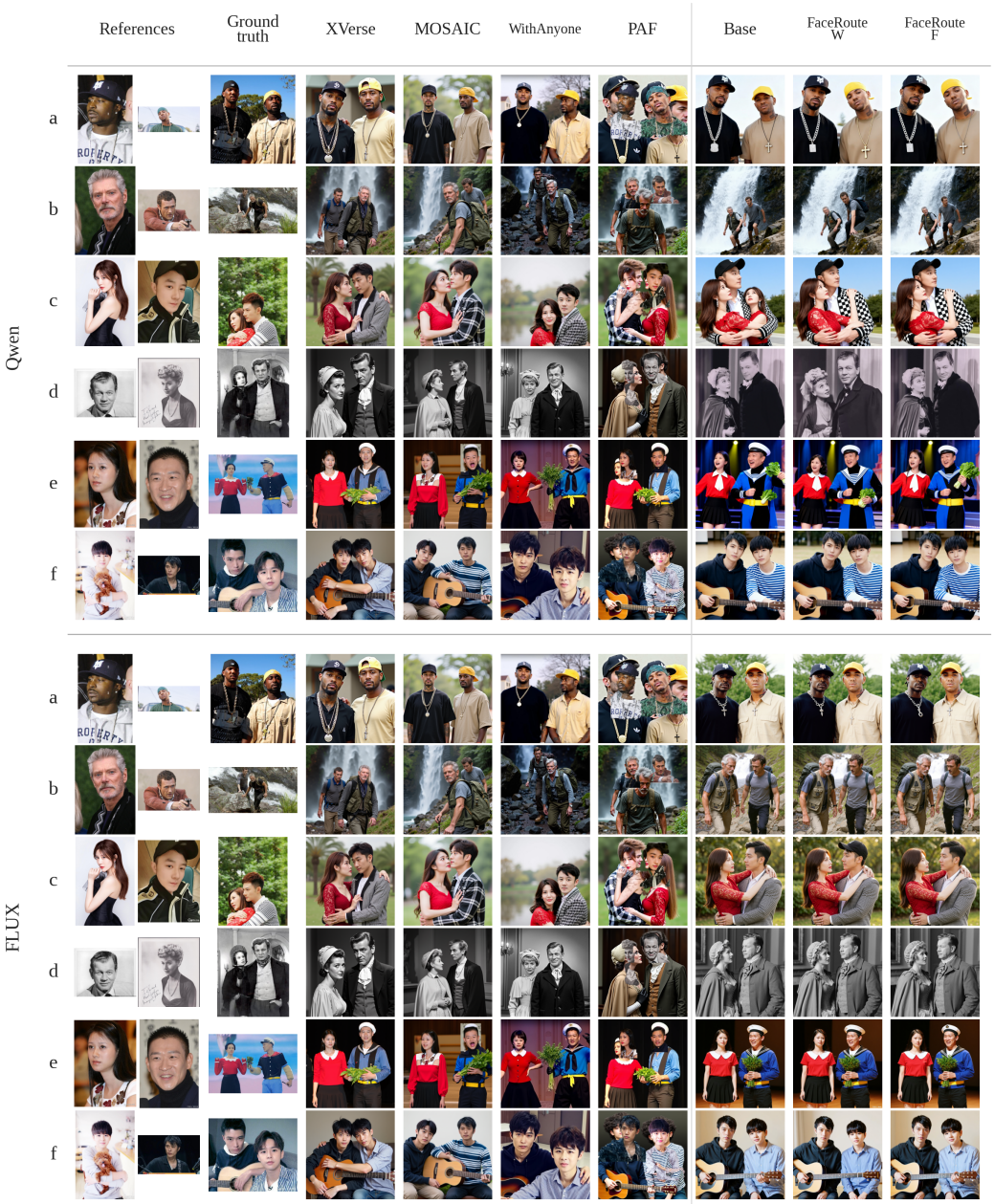

Figure 4: MultiID-Bench two-person examples. Each group shows two identity references, ground truth, and generated outputs from baseline, FaceRoute-W, and FaceRoute-F ($\alpha = 1.5$, Qwen-Image-Edit-2511, 40 steps).

B Data Alignment Audit

The canonical 40-step Qwen and FLUX.2-klein comparisons on the MultiID-Bench two-person (p2) and three-/four-person (p3) splits are aligned by split, prompt, evaluator, resolution, and fixed seed. Table 1 reports completed Qwen and FLUX evaluations on all 194 p2 and 33 p3 inputs. The two historical p3 exclusions lacked `ori.jpg` in evaluation staging, not in the source benchmark: their ground truths were stored as `gt.webp` and `gt.png` and missed by JPG-only staging. Reevaluation uses

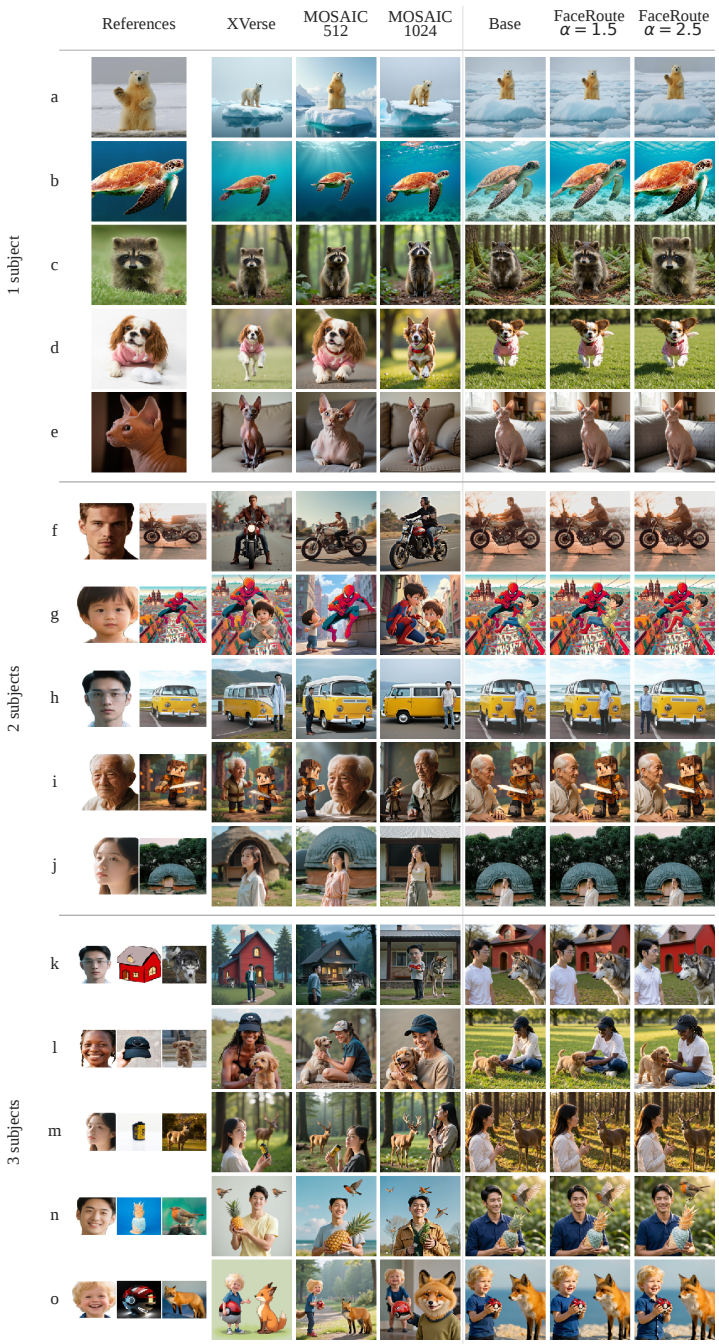

Figure 5: XVerseBench examples. Each group shows subject references and generated outputs from the FLUX.2-KLEIN-BASE-9B baseline and FACEROUTE (mask) variant.

these original files and the existing generated outputs. The key-scope and Whole–Base comparisons use the completed 33-ID P3 cohort from the same original generation arms. CP pair counts remain metric-specific. The Qwen small-face scope comparison uses 49 examples: the original aligned 31 plus 18 supplementary examples, pooled before IQR filtering (Appendix M.4). The Qwen and FLUX small-face tables use the same aligned 31-example source manifest, excluding the corrupted samples `0021` and `0030` and the prompt-invalid `0032` from the original local collection. All full-reference

rows retain 31 evaluator-valid outputs; the FLUX crop baseline retains 30 after validity filtering. These are within-backbone stress tests, not an official benchmark, and no cross-backbone absolute ranking is claimed.

The 4-step p3 baseline, constant all-target FACEROUTE-F routing, and layer-decay rows share 31 valid IDs and have complete identity, CP, Aesthetic, CLIP-I, CLIP-T, and FID summaries. For p2, the original DMDR-4 generation directory and official evaluation summary are traceable and report 171/194 valid examples. We restore its identity, ArcFace, CP, non-match, Aesthetic, and FID fields in Table 16. CLIP-I/T are backfilled separately on exactly these 171 successful IDs, without recomputing the other metrics. Because the baseline and routed successful-ID sets have not been intersected, the p2 comparison remains descriptive rather than paired. A machine-readable audit is kept alongside the draft source.

Evaluator aggregates versus paired statistics. The benchmark evaluator computes summary statistics separately for each method and metric. Its default IQR filter retains scores in $[Q_1 - 1.5\,\mathrm{IQR},\ Q_3 + 1.5\,\mathrm{IQR}]$, where $\mathrm{IQR} = Q_3 - Q_1$. The implementation falls back to robust statistics if filtering would discard more than half the values. Consequently, processed-record counts, metric-valid counts, and retained counts after outlier removal are distinct. Qwen at $\alpha = 2.5$ now retains all 194 two-person records after evaluation retries and a null-handling fix. Undefined values remain null and do not enter numeric IQR filtering; available metrics from the same record still contribute. Processed/Total therefore reports coverage, not complete metric availability or correct person counts. Archived runs retain their recorded populations.

Paired analyses instead intersect successful IDs and use untrimmed per-sample scores, averaging available face-encoder scores as specified by each audit. Metric-specific intersections are retained rather than imputed. The Qwen three-seed CP table uses one common 177-ID intersection across all six generation arms; the Qwen single-seed CP comparison uses its own 187-ID intersection. Their S0 effects therefore differ ($+0.0271$ versus $+0.0263$) without inconsistency. On the completed 33-ID Qwen scope cohort, whole-reference minus face-local reference-similarity aggregates give -0.0006, whereas the untrimmed paired difference is $+0.0147$ (Tables 1 and 2).

The FLUX three-seed study uses 194 identity pairs per seed and ArcFace-only CP on seed-specific intersections of 176, 175, and 174 IDs. It does not use the Qwen common-177 protocol or the main table's three-encoder CP aggregate (Appendix G.2).

C OFFICIAL KEY-SCOPE PAIRED ANALYSIS

For the face-local and whole-reference rows in Table 21, we intersect successful sample IDs and compute an untrimmed per-sample mean over ArcFace, FaceNet, and AdaFace. We report whole-reference minus face-local deltas, paired percentile-bootstrap 95% intervals (10,000 replicates), and two-sided sign-flip tests (100,000 replicates; seed 20260814). The common-ID counts are 194 for P2 and 33 for P3. CP uses 22 Qwen and 27 FLUX P3 pairs with defined scores in both arms. These paired values can differ from the main table's configured method-wise aggregation and are used to judge scope differences rather than compare rounded aggregate means. Table 3 reports the full set.

The only resolved FLUX identity-scope difference is the small p2 advantage for whole-reference routing; FLUX p3 does not distinguish the scopes. Qwen p2 resolves both a whole-reference $\mathrm{sim}_{\mathrm{ref}}$ advantage and a higher CP cost, whereas Qwen p3 does not resolve an identity difference but shows a higher CP cost for whole-reference routing. The data therefore support reference reweighting as the cross-backbone effect and face localization as a backbone- and split-dependent regularizer, not a universal source of stronger identity similarity.

The FLUX three-/four-person paired mean ordering—higher $\mathrm{sim}_{\mathrm{ref}}$, lower $\mathrm{sim}_{\mathrm{GT}}$, and higher CP for whole-reference routing—is descriptive, not three independent pieces of evidence for copying. For a fixed matched face triplet, CP increases algebraically with higher reference similarity and lower ground-truth similarity (Eq. 6); it does not measure background transfer. Its paired $+0.0014$ estimate uses 27 common CP-valid samples, not the difference of arm means with different valid sets. Neither these unresolved contrasts nor the corresponding reward trend establishes that localization becomes more beneficial with additional people.

Table 3: Exact-common-ID paired key-scope analysis at $\alpha = 1.5$. Deltas are whole-reference minus face-local; positive $\Delta\mathrm{M}_{\mathrm{CP}}$ is a cost. Identity comparisons use 194 P2 and 33 P3 pairs; CP counts are metric-specific. A confidence interval crossing zero is treated as unresolved.

Backbone	Split	Metric	n	Δ [95% CI]	p
QWEN-IMAGE-EDIT-2511	p2	$\Delta\mathrm{sim}_{\mathrm{ref}}$	194	$+0.0113$ $[+0.0031, +0.0202]$	0.00775
QWEN-IMAGE-EDIT-2511	p2	$\Delta\mathrm{sim}_{\mathrm{GT}}$	194	$+0.0043$ $[-0.00004, +0.0087]$	0.0574
QWEN-IMAGE-EDIT-2511	p2	$\Delta\mathrm{M}_{\mathrm{CP}}$	187	$+0.0115$ $[+0.0036, +0.0196]$	0.00363
QWEN-IMAGE-EDIT-2511	p3	$\Delta\mathrm{sim}_{\mathrm{ref}}$	33	$+0.0147$ $[-0.0022, +0.0306]$	0.0954
QWEN-IMAGE-EDIT-2511	p3	$\Delta\mathrm{sim}_{\mathrm{GT}}$	33	-0.0027 $[-0.0112, +0.0050]$	0.5180
QWEN-IMAGE-EDIT-2511	p3	$\Delta\mathrm{M}_{\mathrm{CP}}$	22	$+0.0283$ $[+0.0093, +0.0467]$	0.0104
FLUX.2-KLEIN-BASE-9B	p2	$\Delta\mathrm{sim}_{\mathrm{ref}}$	194	$+0.0077$ $[+0.0029, +0.0126]$	0.0020
FLUX.2-KLEIN-BASE-9B	p2	$\Delta\mathrm{sim}_{\mathrm{GT}}$	194	$+0.0074$ $[+0.0035, +0.0114]$	0.00033
FLUX.2-KLEIN-BASE-9B	p2	$\Delta\mathrm{M}_{\mathrm{CP}}$	178	$+0.0028$ $[-0.0016, +0.0074]$	0.2357
FLUX.2-KLEIN-BASE-9B	p3	$\Delta\mathrm{sim}_{\mathrm{ref}}$	33	$+0.0033$ $[-0.0098, +0.0161]$	0.6367
FLUX.2-KLEIN-BASE-9B	p3	$\Delta\mathrm{sim}_{\mathrm{GT}}$	33	-0.0009 $[-0.0100, +0.0076]$	0.8551
FLUX.2-KLEIN-BASE-9B	p3	$\Delta\mathrm{M}_{\mathrm{CP}}$	27	$+0.0014$ $[-0.0072, +0.0112]$	0.7770

D IDENTITY-CONSISTENT MATCHING AUDIT

The released evaluator independently optimizes generated-to-GT and generated-to-reference assignments. A generated face can consequently be paired with GT identity A and reference identity B in one CP triplet. We retain that protocol in the main tables for direct MultiID-Bench comparability and run a stricter reference-anchored audit as a robustness check. In this audit, generated and GT faces are independently assigned one-to-one to the same set of reference identities, and triplets (g_i, t_i, r_i) are then constructed by reference identity. Samples require equal, nonzero generated, GT, and reference face counts. Baseline and $\alpha = 1.5$ use exact common IDs; statistics average ArcFace, FaceNet, and AdaFace within each sample without outlier trimming.

Identity mismatch under independent assignment is empirically nonzero. For baseline/method, FaceNet mismatch rates are 13.5%/12.4% on p2 and 27.1%/30.0% on p3; ArcFace rates are 0.55%/0% and 7.46%/7.46%; AdaFace rates are 2.94%/1.76% and 12.9%/4.29%, respectively. Table 4 shows positive identity and CP changes with intervals above zero after reference anchoring. The directions thus persist on this stricter retained subset, rather than arising only under independent assignment. Because eligibility also changes, the audit does not isolate matching as the sole source of score differences. It neither makes CP a geometry metric nor evaluates identity-to-position binding.

Table 4: Reference-anchored robustness audit for the 40-step Qwen base configuration with $\alpha = 1.5$ versus the unrouted baseline on exact common IDs. Positive $\Delta\mathrm{M}_{\mathrm{CP}}$ is a cost. Intervals are paired-bootstrap 95% CIs (10,000 replicates); p is a two-sided sign-flip test (100,000 replicates; seed 20260812).

Split	Metric	n	Δ [95% CI]	p
p2	$\Delta\mathrm{sim}_{\mathrm{ref}}$	168	$+0.0450$ $[+0.0378, +0.0522]$	$\leq 10^{-5}$
p2	$\Delta\mathrm{sim}_{\mathrm{GT}}$	168	$+0.0256$ $[+0.0200, +0.0312]$	$\leq 10^{-5}$
p2	$\Delta\mathrm{M}_{\mathrm{CP}}$	168	$+0.0263$ $[+0.0197, +0.0331]$	$\leq 10^{-5}$
p3	$\Delta\mathrm{sim}_{\mathrm{ref}}$	21	$+0.0610$ $[+0.0442, +0.0798]$	$\leq 10^{-5}$
p3	$\Delta\mathrm{sim}_{\mathrm{GT}}$	21	$+0.0317$ $[+0.0195, +0.0451]$	6.0×10^{-5}
p3	$\Delta\mathrm{M}_{\mathrm{CP}}$	21	$+0.0420$ $[+0.0253, +0.0592]$	1.9×10^{-4}

E POSITION-PARTITIONED QUERY-SCOPE CONTROL

As a diagnostic, we tested a two-person position-partitioned control (`halves`) that divides target tokens by packed-grid column into $\mathcal{T}_L$ and $\mathcal{T}_R$. Its bias is

$$B_{qk} = \log\alpha\,\big(\mathbb{1}[q \in \mathcal{T}_L,\, k \in \mathcal{F}_1] + \mathbb{1}[q \in \mathcal{T}_R,\, k \in \mathcal{F}_2]\big). \tag{7}$$

Thus each target partition is biased toward one reference face. Table 5 compares this control with all-target FACEROUTE-F at matched strength. All-target FACEROUTE-F gives higher overall matched identity similarities and lower FID, while the position-partitioned control gives lower non-matched similarity and CP. Because evaluation uses optimal matching, this experiment does not establish which method better follows a requested spatial assignment.

Table 5: 4-step LoRA p2 query-scope diagnostic at $\alpha = 2$. The recorded artifact provides identity/copy, Aesthetic, and FID values; ArcFace and CLIP values are not available for every row and are shown as unavailable here. Table 16 lists the available baseline ArcFace and CLIP-I/T values and the $\mathcal{T}_{\text{all}}$ CLIP-I/T values. $\mathcal{T}_{\text{all}}$ denotes all-target queries; the routed row uses FACEROUTE-F.

Method	Valid	sim_{ref}	sim_{GT}	Arc-ref	Arc-GT	M_{CP}	sim_{mis}	Aes.	CLIP-I/T	FID
Baseline	171/194	0.5658	0.4022	–	–	**0.1936**	0.2293	5.6495	–	122.24
`halves`	172/194	0.6465	0.4374	–	–	0.2653	**0.1546**	5.6350	–	121.14
$\mathcal{T}_{\text{all}}$	171/194	**0.6623**	**0.4458**	–	–	0.2739	0.1683	5.6317	–	**118.79**

F REPRODUCIBILITY SCOPE

Table 6 separates the protocol documented here from local integrity records and remaining release requirements. Local manifests are not a public, immutable artifact.

Table 6: Recorded protocol and remaining reproducibility gaps.

Item	Information available
Method	Attention rule, fixed-logit mass relation, SCRFD-to-packed-VAE mapping, scopes, and default strength
Image backbones	Qwen-Image-Edit-2511 and FLUX.2-klein-base-9B; 40 steps, CFG 4, 1024^2, bfloat16
Randomness	Three generation seeds per backbone on the two-person split; other studies use the seeds specified in their protocols
Evaluation	Metric-specific valid sets, IQR-filtered aggregates versus untrimmed paired effects, and bootstrap settings
Video extension	Ten single-person MiniMax-H3 FL2VA inputs, two-pass tracking, seed 42, 124 frames at 24 fps; protocol in Appendix X
Integrity records	Local ledgers retain canonical run IDs and SHA-256 hashes of core code and results; they are not a public release
Runtime	Five-sample H100 image measurements; no controlled peak-memory study
Remaining artifacts	A public code snapshot, complete upstream checkpoint revisions or weight hashes, and a redistributable benchmark manifest are not yet available

Upon acceptance, we plan to release the implementation, evaluation scripts, experiment configurations, and artifact manifests, with pinned code and model revisions. Source-image redistribution remains subject to dataset licenses. The present results should not be read as already accompanied by that release.

F.1 NUMERICAL COPY-PASTE IMPLEMENTATION

For each face encoder ϕ, the evaluator uses $\varepsilon = 10^{-6}$ and

$$\widehat{\phi}_\varepsilon(a) = \frac{\phi(a)}{\|\phi(a)\|_2 + \varepsilon}, \qquad \theta_\varepsilon(a,b) = \arccos\Big(\text{clip}_{[-1,1]}\Big[\widehat{\phi}_\varepsilon(a)^\top \widehat{\phi}_\varepsilon(b)\Big]\Big). \tag{8}$$

The normalized dot product is clipped before $\arccos$. Equation 6 returns zero when $\theta_\varepsilon(t,r) < \varepsilon$ and otherwise divides by the computed $\theta_\varepsilon(t,r)$; this branch is not a $\max(\theta_\varepsilon(t,r),\varepsilon)$ denominator. ArcFace, FaceNet, and AdaFace use the same numerical routine.

G PAIRED STATISTICAL ANALYSIS

For the 40-step Qwen baseline versus FACEROUTE-F with all-target queries ($\mathcal{T}_{\text{all}}$) at $\alpha = 1.5$, we intersected successful IDs before computing the untrimmed per-sample average of ArcFace, FaceNet, and AdaFace. Table 7 reports paired mean changes, percentile 95% bootstrap intervals (10,000 replicates), and two-sided Monte Carlo sign-flip tests (100,000 replicates; seed 20260818). Metric-level n can be below the common ID count when the evaluator did not emit that metric for a sample. These values are intentionally distinct from the main table's method-wise evaluator aggregate so that every paired comparison uses exactly the same successful IDs. For p2, the baseline is the 194-example 40-step Qwen alpha-sweep matrix baseline, not the older 178-example artifact retained for qualitative comparisons.

Table 7: Common-ID paired results for the 40-step Qwen base configuration with FACEROUTE-F, all-target queries ($\mathcal{T}_{\text{all}}$), and $\alpha = 1.5$ versus the unrouted baseline. P3 uses the historical 31-ID cohort. Positive $\Delta\text{M}_{\text{CP}}$ is a cost. CIs are paired bootstrap 95% intervals; p is two-sided sign-flip.

Split	Metric	n	Δ [95% CI]	p
p2	$\Delta\text{sim}_{\text{ref}}$	194	$+0.0451$ $[+0.0384, +0.0521]$	$\leq 10^{-5}$
p2	$\Delta\text{sim}_{\text{GT}}$	194	$+0.0262$ $[+0.0212, +0.0313]$	$\leq 10^{-5}$
p2	$\Delta\text{M}_{\text{CP}}$	187	$+0.0263$ $[+0.0203, +0.0325]$	$\leq 10^{-5}$
p3	$\Delta\text{sim}_{\text{ref}}$	31	$+0.0605$ $[+0.0424, +0.0807]$	$\leq 10^{-5}$
p3	$\Delta\text{sim}_{\text{GT}}$	31	$+0.0277$ $[+0.0177, +0.0390]$	$\leq 10^{-5}$
p3	$\Delta\text{M}_{\text{CP}}$	20	$+0.0465$ $[+0.0315, +0.0628]$	4.0×10^{-5}

G.1 QWEN THREE-SEED TWO-PERSON REPLICATION

The three-seed study uses the full 194-ID p2 split for both methods at every seed. The generation-seed mapping is S0 = 1091359629774735, S1 = 123456789, and S2 = 987654321, shared by the two methods within each seed. S0 uses the same alpha-sweep matrix baseline and $\alpha = 1.5$ run as Table 1; it does not use the older 178-valid qualitative artifact. S1 and S2 use the final seed-fixed runs. For all seeds we recompute exact-common- ID statistics with 10,000 bootstrap replicates and 100,000 sign flips using analysis seed 20260818, independently of the generation seeds. The three-encoder CP results use the same 177-ID intersection on which CP is available for both methods at all three seeds. Table 8 reports the three-encoder means; Table 9 gives the ArcFace-only check.

Table 8: Three-seed replication on the two-person split. Deltas are FACEROUTE-F at $\alpha = 1.5$ minus the seed-matched baseline; brackets are paired-bootstrap 95% CIs. Identity uses 194 pairs per seed; CP uses the same 177-ID intersection across all six arms. Both identity and CP average ArcFace, FaceNet, and AdaFace; positive $\Delta\text{M}_{\text{CP}}$ is a cost. All per-seed sign-flip $p \leq 10^{-5}$.

Seed	$\Delta\text{sim}_{\text{ref}}$	$\Delta\text{sim}_{\text{GT}}$	$\Delta\text{M}_{\text{CP}}$
S0	$+0.0451$ $[+0.0384, +0.0521]$	$+0.0262$ $[+0.0212, +0.0313]$	$+0.0271$ $[+0.0212, +0.0333]$
S1	$+0.0405$ $[+0.0340, +0.0472]$	$+0.0207$ $[+0.0155, +0.0259]$	$+0.0265$ $[+0.0207, +0.0322]$
S2	$+0.0435$ $[+0.0372, +0.0503]$	$+0.0203$ $[+0.0155, +0.0250]$	$+0.0346$ $[+0.0276, +0.0418]$
Mean±SD	$+0.0431 \pm 0.0023$	$+0.0224 \pm 0.0033$	$+0.0294 \pm 0.0045$

Table 10 lists the corresponding per-seed quality metrics, which are descriptive and are not pooled into a paired test.

G.2 FLUX THREE-SEED TWO-PERSON REPLICATION

We repeat the two-person comparison with FLUX.2-klein-base-9B at generation seeds 42, 123, and 2026, using 40 steps, CFG 4, 1024^2 outputs, and $\alpha = 1.5$. Each arm generates all 194 inputs.

Table 9: ArcFace-only three-seed p2 replication. Deltas are FACEROUTE-F minus the seed-matched baseline; positive $\Delta\mathrm{M}_{\mathrm{CP}}$ is a cost.

Seed	$n_{\mathrm{id}}/n_{\mathrm{CP}}$	$\Delta\mathrm{sim}_{\mathrm{ref}}$ [95% CI]	$\Delta\mathrm{sim}_{\mathrm{GT}}$ [95% CI]	$\Delta\mathrm{M}_{\mathrm{CP}}$ [95% CI]
S0	194/180	$+0.0576$ $[+0.0500, +0.0653]$	$+0.0395$ $[+0.0334, +0.0459]$	$+0.0319$ $[+0.0251, +0.0387]$
S1	194/183	$+0.0523$ $[+0.0444, +0.0601]$	$+0.0332$ $[+0.0272, +0.0391]$	$+0.0298$ $[+0.0233, +0.0364]$
S2	194/174	$+0.0602$ $[+0.0528, +0.0681]$	$+0.0354$ $[+0.0301, +0.0408]$	$+0.0422$ $[+0.0350, +0.0495]$
Mean±SD	–	$+0.0567 \pm 0.0040$	$+0.0360 \pm 0.0032$	$+0.0346 \pm 0.0066$

Table 10: Descriptive p2 quality metrics for the three seed-matched runs. These method-wise values are not pooled into an inferential test.

Seed	Method	Aes. ↑	CLIP-I ↑	CLIP-T ↑	FID ↓
S0	Baseline	5.5301	0.7766	0.3310	115.18
S0	FACEROUTE-F	5.5003	0.7768	0.3318	114.92
S1	Baseline	5.5495	0.7772	0.3298	116.44
S1	FACEROUTE-F	5.5067	0.7793	0.3291	115.74
S2	Baseline	5.5074	0.7779	0.3294	116.87
S2	FACEROUTE-F	5.4990	0.7810	0.3296	115.07

For each seed, paired identity effects average the three face encoders on the same 194 IDs; 10,000 paired-bootstrap resamples quantify sample-level uncertainty. The bootstrap analysis uses base RNG seeds 20260919, 20261000, and 20261100, respectively, with deterministic per-metric offsets; these are distinct from generation seeds. Across-seed mean and sample SD summarize three effects, not 582 independent examples.

Table 11: FLUX seed-matched effects, routed minus baseline. Brackets are paired-bootstrap 95% intervals. Identity uses 194 examples per seed; CP here is ArcFace-only and uses seed-specific valid intersections.

Seed	n_{CP}	$\Delta\mathrm{sim}_{\mathrm{ref}}$	$\Delta\mathrm{sim}_{\mathrm{GT}}$	$\Delta\mathrm{CP}_{\mathrm{Arc}}$
42	176	$+0.0393$ $[0.0338, 0.0448]$	$+0.0218$ $[0.0174, 0.0264]$	$+0.0297$ $[0.0250, 0.0344]$
123	175	$+0.0414$ $[0.0360, 0.0475]$	$+0.0241$ $[0.0200, 0.0283]$	$+0.0344$ $[0.0285, 0.0405]$
2026	174	$+0.0417$ $[0.0362, 0.0475]$	$+0.0264$ $[0.0223, 0.0306]$	$+0.0295$ $[0.0239, 0.0352]$
Mean±SD	–	$+0.0408 \pm 0.0014$	$+0.0241 \pm 0.0023$	$+0.0312 \pm 0.0028$

Table 11 and Table 12 report per-seed effects. Both identity measures improve at every seed, with all six intervals above zero. Their across-seed baseline and routed means are $0.4344 \rightarrow 0.4752$ for reference similarity and $0.3497 \rightarrow 0.3738$ for ground-truth similarity. These untrimmed paired results differ from method-wise IQR-filtered aggregates. ArcFace CP also increases at every seed ($0.1423 \rightarrow 0.1735$ on average). Its denominator and encoder differ from the Qwen 177-ID, three-encoder CP study and the main aggregate table; neither CP levels nor effect magnitudes are directly comparable across these protocols.

Each per-seed Aesthetic interval spans zero; this does not establish non-inferiority or absence of quality cost. The replication supports the identity-gain direction across seeds for the tested two-person FLUX protocol. It does not supply multi-seed evidence for the smaller splits, XVerseBench, or video extension.

Table 12: FLUX three-seed quality effects. Mean±SD summarizes the three seed-matched paired changes, not pooled per-image scores.

Metric	Seed 42	Seed 123	Seed 2026	Mean±SD
Aesthetic	-0.0151	-0.0112	-0.0048	-0.0104 ± 0.0052
CLIP-I	$+0.0025$	$+0.0040$	$+0.0019$	$+0.0028 \pm 0.0011$
CLIP-T	-0.0001	-0.0007	$+0.0001$	-0.0003 ± 0.0004

H Position-Binding Scope and GT-Relative Identity Order

The position-partitioned control in Sec. E is a diagnostic for query scope, not a position-aware evaluation, and the primary evaluator uses optimal matching and therefore measures unordered identity-set retention. Because routing raises attention to all reference-face keys at once, a natural concern is whether identity gains also change spatial assignment. We measure correspondence with the ground-truth (GT) identity order. This is not a direct test of instruction-conditioned binding: a prompt can permit compositions other than the GT layout.

Protocol. For each benchmark example we detect faces in the generated image and in the ground-truth image with InsightFace (`buffalo_l`, detection threshold 0.45, $1.5\times$ padded canvas), assign detections to references by maximum-cosine Hungarian matching, and order the assigned identities left to right by bounding-box centre. An example counts as an *order disagreement* when the generated left-to-right identity order differs from the ground-truth order. Examples are scored only when both images resolve every reference exactly once, which makes the comparison paired by construction. We report the McNemar contingency table and a paired bootstrap interval over $10{,}000$ resamples (seed 20260918).

Result. On Qwen-Image-Edit-2511 two-person inputs, all 194 examples are eligible under both arms. The baseline disagrees with GT order in 69 examples (35.6%) and $\mathcal{T}_{\text{all}}$ at $\alpha = 1.5$ in 68 (35.1%); the discordant cells are small and nearly symmetric (4 baseline-only, 3 routed-only, 65 both, 122 neither), giving a paired change of -0.0052 with a 95% interval of $[-0.0309, +0.0206]$. On three-/four-person inputs, 26 of 33 examples are eligible; the baseline disagrees in $21/26$ (80.8%) and routing in $18/26$ (69.2%), a paired change of -0.1154 with interval $[-0.2692, 0.0000]$, which touches zero and is therefore unresolved. All three discordant examples on this split have agreement only under routing.

Interpretation. The two-person disagreement-rate change is unresolved: the interval permits either a reduction of 3.09 percentage points or an increase of 2.06 points. The smaller split also does not resolve a change. All 194 two-person examples satisfy the audit's face-count and assignment criteria under both arms, but this eligibility does not establish correct identity-to-position binding. The results show that improved identity-set retention can coexist with different GT-relative arrangements. When the prompt does not specify the GT order, these are not necessarily generation errors. The results neither establish non-inferiority in order disagreement nor show that routing solves position binding; that claim would require explicit identity-to-position instructions and a corresponding evaluator.

I Controlled Inference Throughput

We measure denoising-loop throughput for the 40-step Qwen base configuration on one H100 80GB GPU using p2 inputs with two references, 1024×1024 outputs, batch size one, 40 steps, true CFG 4, and the fixed seed used elsewhere. The three variants process the same six examples sequentially on the same device. We exclude the first example of each variant as warm-up; for FlexAttention this also excludes its first-use `torch.compile` cost. Each reported denoising step includes the two transformer forward passes required by true CFG. Timing covers the denoising loop only and excludes model loading, reference preprocessing, and VAE decoding.

Dense routing is not negligible in this implementation: it retains 54.2% of baseline throughput. FlexAttention is $1.56\times$ faster than dense routing and recovers throughput to 84.9% of baseline, while remaining 15.1% slower than the unmodified model. This is a single-device runtime measurement

Table 13: Controlled true-CFG inference throughput. A denoising step/s includes two transformer forwards. Values average five matched post-warm-up examples.

Method	Seconds/image ↓	Steps/s ↑	Rel. throughput ↑
Baseline	65.2	0.614	100.0%
$\mathcal{T}_{\text{all}}$ $\alpha = 1.5$ (dense)	120.2	0.333	54.2%
$\mathcal{T}_{\text{all}}$ $\alpha = 1.5$ + FlexAttention	76.8	0.521	84.9%

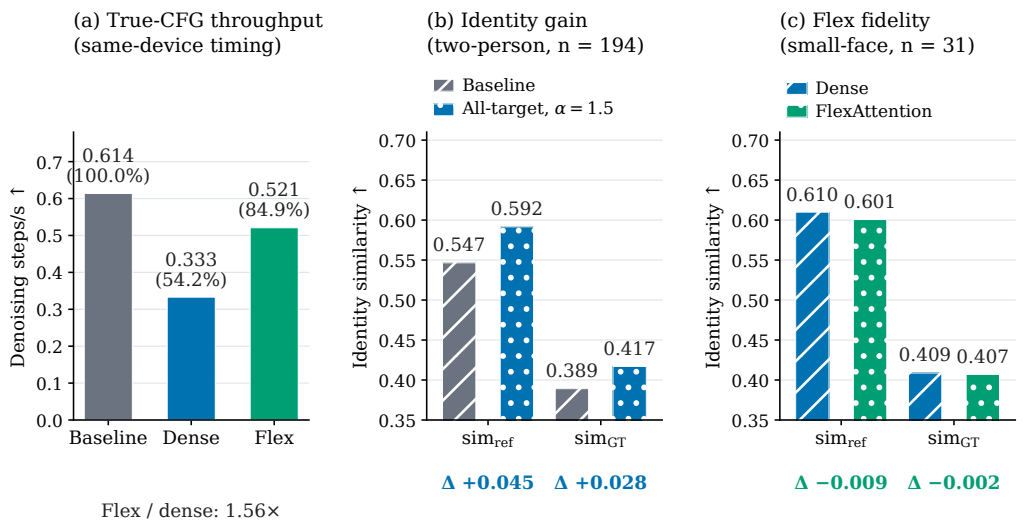

Figure 6: Runtime and identity comparison. (a) Controlled p2 true-CFG timing on one H100; percentages are throughput relative to baseline, and one denoising step includes two transformer forwards. (b) Identity improvement of dense $\mathcal{T}_{\text{all}}$ at $\alpha = 1.5$ over baseline on the complete p2 evaluation ($n = 194$). (c) Dense-versus-FlexAttention fidelity on the separate matched 31-example `smallface_flux` manifest. Panels (b) and (c) use different manifests and isolate different comparisons; their absolute SIM values should not be interpreted as a three-method same-set ranking.

rather than a general efficiency claim; peak memory is not measured. Figure 6 shows the timing and identity panels.

J FlexAttention Fidelity Check

We also test whether a block-sparse FlexAttention implementation changes the behavior of the final routing configuration. This is an implementation comparison, rather than an ablation of the routing method: both rows use the 40-step Qwen base configuration, all-target face-key routing with $\alpha = 1.5$, no reference isolation, and the same fixed generator seed. We evaluate the complete 31-example converted small-face manifest (whose artifact directory is named `smallface_flux`). Table 14 shows that FlexAttention is close to, but not exactly interchangeable with, the dense implementation. The average reference-identity score decreases by 0.0092, while the other aggregate differences are also small. Because iterative denoising can amplify small attention-level numerical differences, we do not describe the sparse implementation as output-identical or lossless. Together with the controlled timing result in Table 13, these results support treating FlexAttention as a measured acceleration implementation: it is $1.56\times$ faster than dense routing in that protocol, with a modest fidelity trade-off that should be revalidated when changing the inference stack.

Table 14: Dense versus FlexAttention implementation check for the final all-target face-key routing configuration evaluated with the 40-step Qwen base configuration on the 31-example converted small-face manifest. Both implementations use the same seed and inference settings. This table is descriptive: it reports one paired generation per example, not a multi-seed significance test.

Implementation	$\mathrm{sim}_{\mathrm{ref}} \uparrow$	$\mathrm{sim}_{\mathrm{GT}} \uparrow$	$\mathrm{M}_{\mathrm{CP}} \downarrow$	Aes. $\uparrow$
Dense attention	0.6099	0.4088	0.2279	5.2177
FlexAttention	0.6007	0.4070	0.2333	5.2016

Table 15: Full FLUX.2 results for the unrouted baseline and FACEROUTE-F. The baseline and $\alpha = 1.5$ rows are the FLUX.2 Base and FACEROUTE-F 1.5 rows of Table 1; $\alpha = 2.0$ is an additional strength. Each row uses the same 40-step, CFG-4, seed-42 protocol, with 194 two-person and 33 three-/four-person records. Identity scores pool three encoders before metric-wise IQR filtering; Arc-ref and Arc-GT use ArcFace alone. The missing two-person $\mathrm{sim}_{\mathrm{mis}}$ summary at $\alpha = 1.5$ is not imputed. Here $\mathrm{sim}_{\mathrm{mis}}$ measures non-matched reference similarity, not the GT-based Bld in Table 1. Values are descriptive method-wise aggregates.

(a) Identity diagnostics

Split	α	n_{proc}	$\mathrm{sim}_{\mathrm{ref}}$	$\mathrm{sim}_{\mathrm{GT}}$	Arc-ref	Arc-GT	$\mathrm{sim}_{\mathrm{mis}}$
2-person	1 (baseline)	194	0.4388	0.3465	0.4321	0.3062	0.0739
	1.5	194	0.4789	0.3684	0.4891	0.3433	–
	2.0	194	0.5069	0.3867	0.5267	0.3670	0.0750
3–4-person	1 (baseline)	33	0.3626	0.3044	0.2847	0.2045	0.0220
	1.5	33	0.4215	0.3391	0.3608	0.2561	0.0227
	2.0	33	0.4552	0.3580	0.4081	0.2824	0.0328

(b) Reference preference and global quality

Split	α	M_{CP}	Aesthetic	CLIP-I	CLIP-T	FID
2-person	1 (baseline)	0.1044	5.8342	0.7527	0.3365	118.35
	1.5	0.1204	5.8098	0.7553	0.3359	117.73
	2.0	0.1398	5.8137	0.7584	0.3353	118.16
3–4-person	1 (baseline)	0.0759	6.4927	0.7128	0.3388	195.34
	1.5	0.1014	6.4741	0.7157	0.3402	195.44
	2.0	0.1093	6.4094	0.7208	0.3378	191.97

K QUALITATIVE RESULTS AND FAILURE CASES

The two-person qualitative comparison is presented in Figure 1. Figure 10 shows selected small-face examples. These are illustrations, not estimates of success or failure frequency. Quantitative examples of worsened competing-identity similarity and face collapse are reported in Appendix O; they demonstrate that individual failures remain despite the aggregate identity gains.

L SUPPORTING TABLES AND FIGURES FROM THE MAIN TEXT

Full FLUX metrics. Table 15 reports additional metrics for the FLUX.2 baseline and FACEROUTE-F. Its baseline and $\alpha = 1.5$ rows are the FLUX.2 Base and FACEROUTE-F 1.5 runs of Table 1, with unchanged values and evaluator aggregation; $\alpha = 2.0$ is an additional strength not shown there. The two panels separate identity diagnostics from global-quality proxies for readability.

Four-step LoRA configuration. As a supplementary transfer experiment, we evaluate QWEN-IMAGE-EDIT-2511 with its few-step LoRA scheduler branch using four denoising steps, true CFG 1, 1024×1024 resolution, and one fixed seed per comparison. These results are reported separately because their absolute scores are not directly comparable to those of the 40-step base configuration.

Table 16: 4-step LoRA results, reported separately from the 40-step stack. All routed rows use $\mathcal{T}_{\text{all}}$ with no reference isolation. The p3 rows share 31 valid IDs. The p2 baseline is the traceable official DMDR-4 evaluation; its CLIP-I/T values are backfilled on the same 171 successful IDs. The two p2 rows are descriptive and non-paired because their successful-ID sets have not been intersected.

(a) Identity diagnostics

Split	Method	Valid	$\text{sim}_{\text{ref}} \uparrow$	$\text{sim}_{\text{GT}} \uparrow$	Arc-ref $\uparrow$	Arc-GT $\uparrow$
p2	Baseline	171/194	0.5658	0.4022	0.5533	0.3799
p2	$\alpha = 2.0$	171/194	0.6623	0.4458	–	–
p3	Baseline	31/33	0.6540	0.4294	0.5694	0.3674
p3	$\alpha = 2.0$	31/33	0.7010	0.4529	0.6628	0.4063
p3	Layer $2 \to 1.5$	31/33	0.6860	0.4485	0.6231	0.3947

(b) Reference preference and global quality

Split	Method	$\text{M}_{\text{CP}} \downarrow$	$\text{sim}_{\text{mis}} \downarrow$	Aes. $\uparrow$	CLIP-I $\uparrow$	CLIP-T $\uparrow$	FID $\downarrow$
p2	Baseline	0.1936	0.2293	5.6495	0.7773	0.3254	122.24
p2	$\alpha = 2.0$	0.2739	0.1683	5.6317	0.7846	0.3254	118.79
p3	Baseline	0.2644	0.1204	5.7323	0.7400	0.3170	194.72
p3	$\alpha = 2.0$	0.3410	0.1141	5.6729	0.7420	0.3139	208.37
p3	Layer $2 \to 1.5$	0.3106	0.1126	5.6708	0.7365	0.3130	200.01

Table 17: Retrospective cross-split alpha diagnostic for the 40-step Qwen base configuration on the historical 31-ID p3 cohort, not the completed 33-record Table 1 cohort. The criterion was formulated after completing the sweep; this is a consistency check, not held-out hyperparameter validation.

Selection split	Chosen α	Test split	$\Delta\text{sim}_{\text{ref}}$	Δ FID
p2	1.5	p3	$+0.0824$	-2.05

Scheduling ablation. Table 23 reports the layer and step schedules referenced in Section 7, separating identity diagnostics from quality metrics on both splits.

M Ablation Studies

M.1 Routing Strength

Figure 7 and Table 18 summarize the complete constant-strength sweep. Identity similarity generally increases with α on both splits. M_{CP} rises at the same time, exposing a retention–preference trade-off. The observed FID is non-monotonic; without FID uncertainty we use it only as a descriptive warning, most notably for the larger p3 value at $\alpha = 2.5$, rather than as evidence that small differences rank image quality.

M.2 Retrospective Split-Half Sensitivity of Routing Strength

The default $\alpha = 1.5$ was chosen after inspecting the complete sweep. We therefore use split-half resampling to describe its within-benchmark stability, not as independent validation. The selection rule chooses the smallest swept α whose paired $\Delta\text{sim}_{\text{ref}}$ reaches 0.04. Both the rule and threshold are retrospective. We draw 1000 random balanced splits of benchmark examples (seed 20260918), select on one half, and measure the selected arm on the other. Scoring both fold orders gives 2000 measurements, which reuse the same examples and are not independent trials.

On the two-person split ($n = 191$ common examples), the rule selects $\alpha = 1.5$ in 93.9% of folds and $\alpha = 1.75$ in 6.1%. The evaluation-half gains average $+0.0460$ in sim_{ref} (SD 0.0052, minimum $+0.0335$) and $+0.0266$ in sim_{GT} (SD 0.0032, minimum $+0.0175$). Both gains are positive in every fold. Their proximity to the full-split gains ($+0.0453$ and $+0.0263$) describes stability under this resampling scheme; it does not rule out selection bias from inspecting the benchmark before specifying the rule.

Table 18: Constant-α sweep on QWEN-IMAGE-EDIT-2511 with 40 steps. All rows use $\mathcal{T}_{\text{all}}$ and no reference isolation. For each of p2 and p3, two panels report the same complete evaluator metric set: three-backend identity means, ArcFace reference and GT scores, M_{CP}, non-match similarity, Aesthetic, CLIP-I, CLIP-T, and FID. This archived sweep retains its original evaluated sets (31 p3 IDs); completed Table 1 rows are not substituted at selected strengths. Bold marks the best method-wise aggregate in each panel; because p2 valid sets differ across α, it is not a paired ranking.

(a) MultiID-Bench p2: identity diagnostics

α	Valid	sim_{ref}	sim_{GT}	Arc-ref	Arc-GT	sim_{mis}
1.00	194/194	0.5466	0.3892	0.5257	0.3630	0.1427
1.25	193/194	0.5741	0.4072	0.5645	0.3935	0.1389
1.50	194/194	0.5916	0.4170	0.5881	0.4057	0.1406
1.75	193/194	0.6094	0.4228	0.6065	0.4165	0.1394
2.00	193/194	0.6207	0.4254	0.6234	0.4211	0.1390
2.50	191/194	**0.6352**	**0.4356**	**0.6448**	**0.4325**	**0.1354**

(b) MultiID-Bench p2: reference preference and global quality

α	M_{CP}	Aes.	CLIP-I	CLIP-T	FID
1.00	**0.1848**	**5.5301**	0.7766	0.3310	115.18
1.25	0.2037	5.5113	0.7774	0.3306	117.10
1.50	0.2129	5.5003	0.7768	0.3318	114.92
1.75	0.2253	5.4931	0.7792	0.3318	116.54
2.00	0.2379	5.5010	0.7771	**0.3322**	116.18
2.50	0.2475	5.4806	**0.7792**	0.3298	116.37

(c) MultiID-Bench p3: identity diagnostics

α	Valid	sim_{ref}	sim_{GT}	Arc-ref	Arc-GT	sim_{mis}
1.00	31/33	0.5353	0.3787	0.4635	0.3089	0.1079
1.25	31/33	0.5950	0.3956	0.5484	0.3368	0.1016
1.50	31/33	0.6177	0.4064	0.5499	0.3506	0.1021
1.75	31/33	0.6269	0.4095	0.5751	0.3553	0.1070
2.00	31/33	0.6345	0.4090	0.5876	0.3598	0.1067
2.50	31/33	**0.6640**	**0.4178**	**0.6204**	**0.3726**	**0.1014**

(d) MultiID-Bench p3: reference preference and global quality

α	M_{CP}	Aes.	CLIP-I	CLIP-T	FID
1.00	**0.2046**	**5.5431**	0.7649	**0.3191**	186.92
1.25	0.2347	5.4506	0.7640	0.3167	185.33
1.50	0.2497	5.4435	**0.7692**	0.3159	184.87
1.75	0.2643	5.4828	0.7657	0.3181	185.91
2.00	0.2761	5.3730	0.7613	0.3163	187.39
2.50	0.3133	5.4177	0.7671	0.3147	194.85

On the smaller three-/four-person split ($n = 31$), selections vary more: $\alpha = 1.25$ accounts for 47.2% of folds and $\alpha = 1.5$ for 51.5%. Evaluation-half gains average $+0.0510$ in sim_{ref} and $+0.0227$ in sim_{GT}, again positive in every fold. Applying the two-person default also gives an observed FID change of -2.05 on this split (Table 17), reported descriptively rather than used for selection. These repeated measurements support the direction of the within-benchmark effect, but do not identify a universally optimal strength or replace validation on new data.

M.3 PER-SAMPLE IDENTITY–COPY COUPLING

Figure 8 compares paired per-sample changes from the same-split baseline to $\mathcal{T}_{\text{all}}$ with $\alpha = 1.5$. On p2 ($n = 187$ samples with all three metrics available), $\Delta\text{sim}_{\text{ref}}$ is strongly associated with $\Delta\text{M}_{\text{CP}}$ (Pearson $r = 0.662$), whereas $\Delta\text{sim}_{\text{GT}}$ is nearly uncorrelated with it ($r = -0.081$). The mean changes are $+0.0444$, $+0.0259$, and $+0.0263$ for sim_{ref}, sim_{GT}, and M_{CP}, respectively. Thus, the gain against the benchmark ground truth does not simply scale with increased reference likeness,

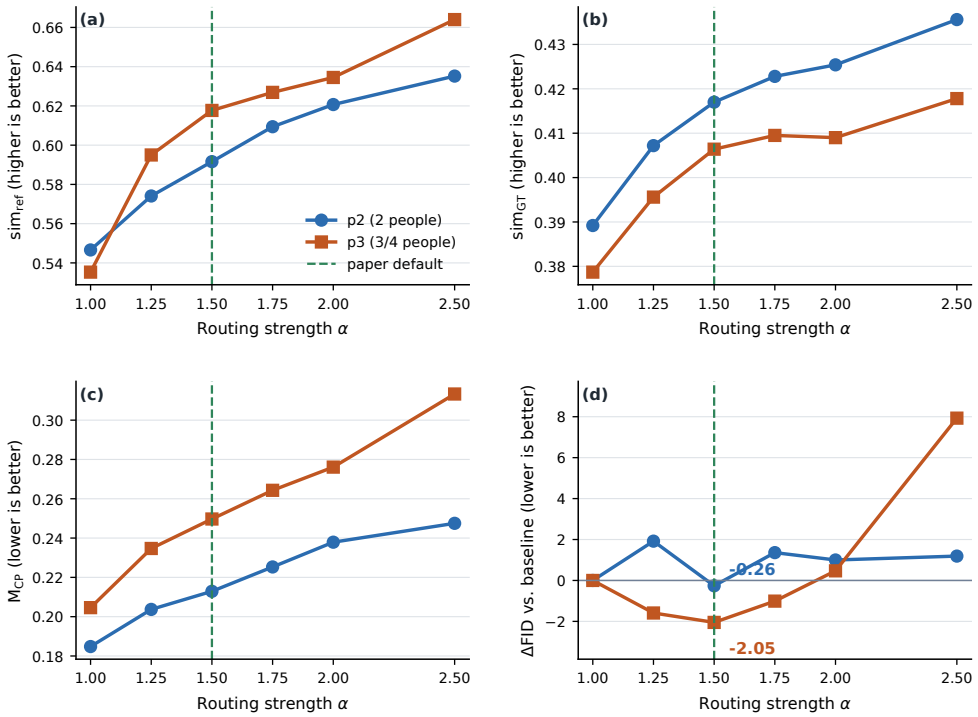

Figure 7: Constant-strength sweep on the 40-step QWEN-IMAGE-EDIT-2511 stack. $\mathrm{sim}_{\mathrm{ref}}$ and $\mathrm{sim}_{\mathrm{GT}}$ generally rise with α, while CP reference preference also increases. FID is plotted as a signed change from each split's baseline; differences are interpreted within each split. The dashed line marks the paper default $\alpha = 1.5$.

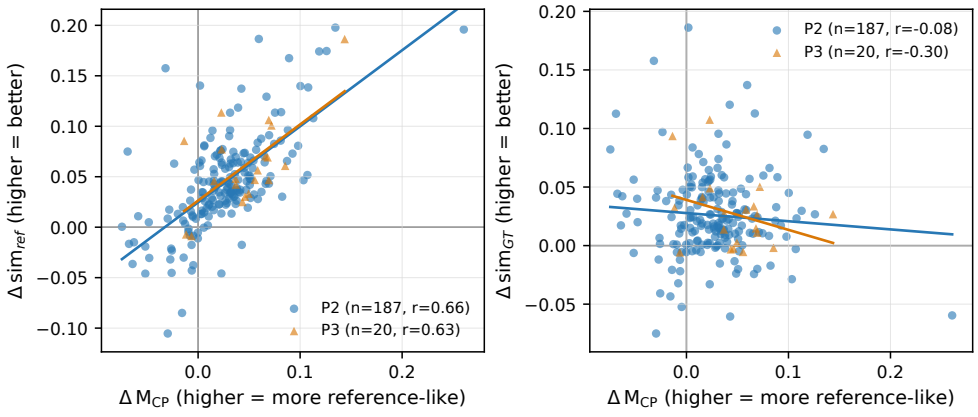

Figure 8: Paired per-sample changes for $\mathcal{T}_{\mathrm{all}}$ at $\alpha = 1.5$ relative to the same-split unrouted 40-step Qwen baseline. Lines are ordinary least-squares fits and r is Pearson correlation. Reference similarity and M_{CP} rise together (left), while the ground-truth identity gain is not positively proportional to the reference-preference increase (right). Because M_{CP} is constructed from overlapping face-embedding angles (Eq. 6), this plot characterizes the observed trade-off and does not establish causality.

even though stronger reference similarity and the copy metric are coupled. The p3 subset shows the same positive $\Delta\mathrm{sim}_{\mathrm{ref}}$–$\Delta\mathrm{M}_{\mathrm{CP}}$ association ($r = 0.631$), but its $\Delta\mathrm{sim}_{\mathrm{GT}}$ correlation ($r = -0.305$) is descriptive only because just 20 samples have every required metric. These correlations should not be read as causal because Eq. 6 shares embedding quantities with both identity measures.

Table 19: Small-face stress test on QWEN-IMAGE-EDIT-2511 at 40 steps using our local, non-official two-person benchmark, not MultiID-Bench p3. Every row uses the same aligned 31-example source manifest, prompts, and fixed seed. Source IDs `0021`, `0030`, and prompt-invalid `0032` are excluded. This is the aligned cross-backbone subset; the main-text Qwen analysis uses the larger 49-example cohort (Figure 3). Identity scores pool available encoder values without IQR filtering. *Face keys* is FACEROUTE-F: only face-local reference keys receive an increased prior from all target queries. *Whole-ref keys* is FACEROUTE-W and biases all keys in each natural full reference. Crop rows replace the model input with resized face crops. Bold marks the better of the two routing scopes on natural full references; arrows indicate metric direction.

(a) Identity diagnostics

Reference input / biased keys	Valid	$\mathrm{sim}_{\mathrm{ref}}\uparrow$	$\mathrm{sim}_{\mathrm{GT}}\uparrow$	Arc-ref ↑	Arc-GT ↑
Full / none (baseline)	31	0.5453	0.3825	0.4598	0.2837
Full / face keys (FACEROUTE-F)	31	**0.6099**	**0.4088**	**0.5331**	0.3164
Full / whole-ref keys	31	0.6016	0.4053	0.5266	**0.3214**
Crop / none	31	0.2077	0.1958	0.1132	0.0997
Crop / all crop keys	31	0.2545	0.2197	0.1719	0.1267

(b) Reference preference and global quality

Reference input / biased keys	$\mathrm{M}_{\mathrm{CP}}\downarrow$	Aes. ↑	CLIP-I ↑	CLIP-T ↑
Full / none (baseline)	0.1875	5.2621	0.7945	0.3368
Full / face keys (FACEROUTE-F)	**0.2279**	**5.2177**	0.7948	0.3357
Full / whole-ref keys	0.2351	5.1764	**0.7970**	**0.3363**
Crop / none	0.0321	4.8337	0.7583	0.3302
Crop / all crop keys	0.0597	4.8231	0.7656	0.3285

Table 20: Small-face stress test on FLUX.2-KLEIN-BASE-9B using the same aligned 31-example source manifest as Table 19. All rows use the same fixed seed and the same unfiltered aggregation as Table 19. The full-reference rows retain all 31 examples; the crop baseline has 30 outputs after evaluator validity filtering. Source IDs `0021`, `0030`, and prompt-invalid `0032` are excluded. Bold compares only the two routing scopes on natural full references; absolute scores are not compared across backbones.

Reference input / biased keys	Valid	$\mathrm{sim}_{\mathrm{ref}}\uparrow$	$\mathrm{sim}_{\mathrm{GT}}\uparrow$	$\mathrm{M}_{\mathrm{CP}}\downarrow$	Aes. ↑	CLIP-I ↑	CLIP-T ↑
Full / none (baseline)	31	0.3000	0.2660	0.0283	5.8498	0.7585	0.3308
Full / face keys (FACEROUTE-F)	31	**0.3254**	**0.2794**	**0.0336**	**5.8417**	**0.7654**	0.3295
Full / whole-ref keys	31	0.3208	0.2719	0.0397	5.7571	0.7630	**0.3324**
Crop / none	30	0.3017	0.2603	0.0494	5.5292	0.7600	0.3351

M.4 THE 49-EXAMPLE SMALL-FACE SUMMARY

The expanded Qwen stress test in Figure 3 contains 49 examples with natural full references: the aligned 31-example set and 18 supplementary examples. Both routing scopes use all 49 examples. The stress test follows the construction methodology of MultiID-Bench (Xu et al., 2026). For each method and metric, we pool the available ArcFace, FaceNet, and AdaFace scores across all 49 examples, apply the evaluator's $1.5\times$IQR filter once to the pooled values, and average the retained scores. Each identity metric has 147 encoder scores before filtering. For $\mathrm{sim}_{\mathrm{ref}}/\mathrm{sim}_{\mathrm{GT}}$, FACEROUTE-W retains 147/146 scores and FACEROUTE-F retains 145/146. Ordinary P2 values in the figure use the same pooled, metric-wise IQR aggregation rule. The 31-example cross-backbone controls in Tables 19 and 20 retain their simple means of available encoder scores without IQR filtering.

For FACEROUTE-W and FACEROUTE-F, respectively, the 49-example scores are $0.6421304507/0.6507770095$ in $\mathrm{sim}_{\mathrm{ref}}$ and $0.3996590532/0.4020397444$ in $\mathrm{sim}_{\mathrm{GT}}$. Differences are computed before display rounding. Face area is computed independently: each detector-valid box area is divided by its VAE-aligned reference-canvas area, then averaged over detected reference images. This yields 1.6276% for 96 of 98 small-face references and 16.3481% for 381 of 388 ordinary P2 references. Undetected references are counted but excluded from the area

Table 21: Matched key-scope control on official MultiID-Bench. FACEROUTE-F biases detected face keys, while FACEROUTE-W biases every key in each natural full reference; both use all target queries and $\alpha = 1.5$. Bold marks the better aggregate within each scope pair and does not imply significance. Both scopes use 194 p2 and 33 p3 inputs, with method-wise pooled, IQR-filtered scores as in Table 1. Paired confidence intervals are reported in Appendix C.

(a) Identity and reference preference

Backbone	Split	Key scope	Valid	$\mathrm{sim}_{\mathrm{ref}} \uparrow$	$\mathrm{sim}_{\mathrm{GT}} \uparrow$	$\mathrm{M}_{\mathrm{CP}} \downarrow$
QWEN-IMAGE-EDIT-2511	p2	FACEROUTE-F	194/194	0.5916	0.4170	**0.2129**
QWEN-IMAGE-EDIT-2511	p2	FACEROUTE-W	194/194	**0.6056**	**0.4214**	0.2289
QWEN-IMAGE-EDIT-2511	p3	FACEROUTE-F	33/33	**0.6057**	**0.4049**	**0.2354**
QWEN-IMAGE-EDIT-2511	p3	FACEROUTE-W	33/33	0.6051	0.4021	0.2550
FLUX.2-KLEIN-BASE-9B	p2	FACEROUTE-F	194/194	0.4789	0.3684	**0.1204**
FLUX.2-KLEIN-BASE-9B	p2	FACEROUTE-W	194/194	**0.4866**	**0.3766**	0.1236
FLUX.2-KLEIN-BASE-9B	p3	FACEROUTE-F	33/33	0.4215	**0.3391**	0.1014
FLUX.2-KLEIN-BASE-9B	p3	FACEROUTE-W	33/33	**0.4247**	0.3383	**0.0981**

(b) Global quality

Backbone	Split	Key scope	Aes. ↑	CLIP-I ↑	CLIP-T ↑	FID ↓
QWEN-IMAGE-EDIT-2511	p2	FACEROUTE-F	**5.5003**	0.7768	**0.3318**	114.92
QWEN-IMAGE-EDIT-2511	p2	FACEROUTE-W	5.4815	**0.7804**	0.3288	**114.40**
QWEN-IMAGE-EDIT-2511	p3	FACEROUTE-F	5.4366	**0.7748**	0.3180	**183.74**
QWEN-IMAGE-EDIT-2511	p3	FACEROUTE-W	**5.4556**	0.7560	**0.3207**	186.82
FLUX.2-KLEIN-BASE-9B	p2	FACEROUTE-F	**5.8098**	0.7553	**0.3359**	**117.73**
FLUX.2-KLEIN-BASE-9B	p2	FACEROUTE-W	5.7901	**0.7611**	0.3358	118.95
FLUX.2-KLEIN-BASE-9B	p3	FACEROUTE-F	**6.4741**	0.7157	0.3402	195.44
FLUX.2-KLEIN-BASE-9B	p3	FACEROUTE-W	6.4198	**0.7186**	**0.3404**	**192.73**

Table 22: Reference-isolation ablation on official MultiID-Bench p3 with the 40-step QWEN-IMAGE-EDIT-2511 base configuration, $\mathcal{T}_{\mathrm{all}}$, $\alpha = 2$, and the same 31 valid IDs. *None (routing on)* applies routing without reference–reference masking and is not the baseline; *Face-only* blocks only cross-reference face-token interactions; *Full* blocks all cross-reference token interactions.

(a) Identity diagnostics

Isolation	$\mathrm{sim}_{\mathrm{ref}}$	$\mathrm{sim}_{\mathrm{GT}}$	Arc-ref	Arc-GT	$\mathrm{sim}_{\mathrm{mis}}$
None (routing on)	**0.6345**	**0.4090**	**0.5876**	**0.3598**	**0.1067**
Face-only	0.5770	0.3902	0.5246	0.3242	0.1292
Full	0.5164	0.3557	0.4579	0.2912	0.1536

(b) Reference preference and global quality

Isolation	M_{CP}	Aes.	CLIP-I	CLIP-T	FID
None (routing on)	0.2761	**5.3730**	**0.7613**	0.3163	**187.39**
Face-only	0.2294	5.3417	0.7594	0.3143	191.28
Full	**0.2006**	5.3669	0.7534	**0.3208**	189.48

mean. Box area is not a measure of facial clarity or an exact packed-token fraction. The cross-dataset comparison also varies content and is not a controlled face-scale intervention.

M.5 REFERENCE-BACKGROUND TRANSFER AUDIT

We randomly sampled 80 eligible P2 IDs (sampling seed 20260924) with both Qwen routing outputs and included the 31 P3 IDs in the aligned scope set. Both outputs for each ID use the same sample prompt and full references; FACEROUTE-W and FACEROUTE-F use $\alpha = 1.5$. One rater viewed the prompt, references, and an anonymized, randomized A/B pair, then chose which output more

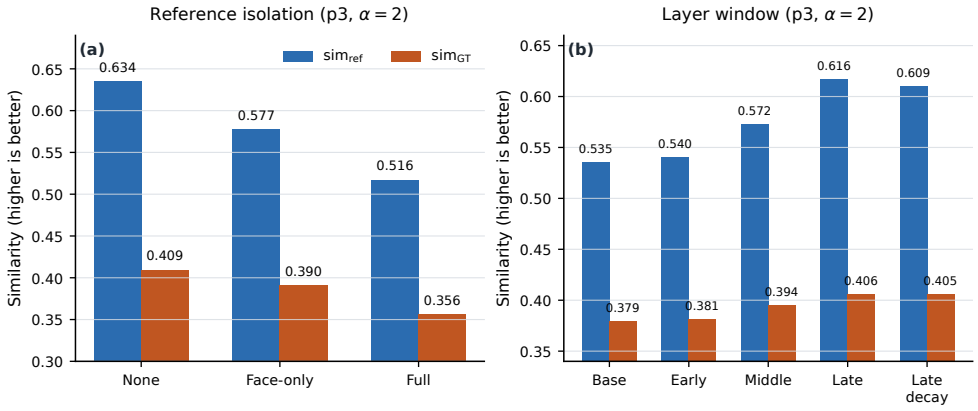

Figure 9: Core routing ablations on the 40-step p3 setting. (a) Hard reference isolation reduces both $\mathrm{sim}_{\mathrm{ref}}$ and $\mathrm{sim}_{\mathrm{GT}}$ under all-target routing. No isolation means routing remains active without a cross-reference mask. (b) Identity retention strengthens when routing moves from early to later transformer blocks; late decay preserves most of this gain.

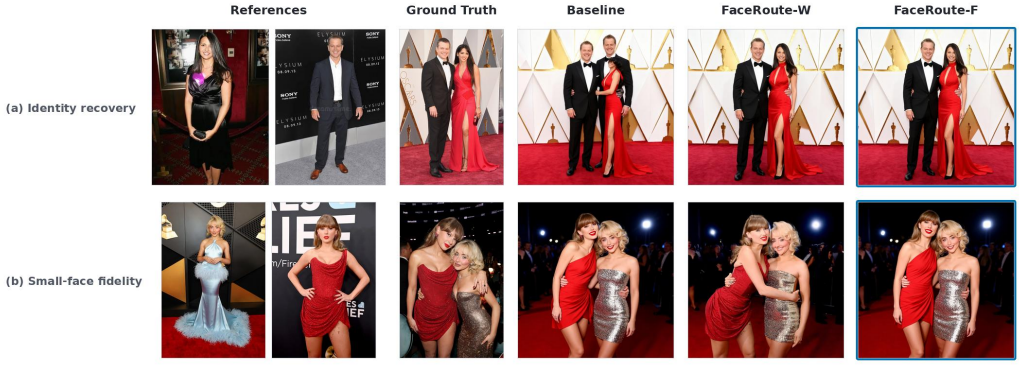

Figure 10: Selected qualitative comparisons from the 49-example Qwen small-face stress test using identical seeds and the 40-step setting. Each row shows the two identity references, ground truth, baseline, FACEROUTE-W at $\alpha = 1.5$, and FACEROUTE-F (all-target queries) at $\alpha = 1.5$. Blue outlines mark the FACEROUTE-F outputs discussed in the text. The examples are selected for visual inspection only; quantitative scope comparisons use all 49 examples (Figure 3).

visibly transferred reference-background elements not requested by the prompt. People, clothing, and general image quality were excluded from this judgment. All 111 pairs received one rating.

On P2, the counts were 9 for FACEROUTE-W more transfer, 1 for FACEROUTE-F more transfer, 2 for both at similar levels, 3 for neither, and 65 for no discernible direction. The corresponding P3 counts were 5, 1, 1, 0, and 24. Thus, 14 of the 16 directionally resolved pairs were rated as showing more transfer under FACEROUTE-W; the remaining 95 ratings retain their separate categories (3 both, 3 neither, 89 undecidable).

M.6 FACE-LOCAL KEYS VERSUS CROPPED INPUTS AND WHOLE-REFERENCE KEYS

Table 19 tests which part of the routing design produces the identity gain on the aligned 31-example small-face subset. The baseline consumes the natural full references without routing. Our setting retains those references and applies an increased prior from every target query only to the detected face keys (all-target queries, $\mathcal{T}_{\mathrm{all}}$, $\alpha = 1.5$). The *whole-reference* control keeps the same target-query scope and strength but expands the biased key set from face tokens to all reference-image tokens.

Table 23: Scheduling ablation on QWEN-IMAGE-EDIT-2511 with 40 steps. Constant and single-axis rows start at $\alpha = 2$; joint applies both a layer and a step $2 \to 1$ decay multiplicatively. The four panels preserve all metrics while separating benchmark subsets and metric groups.

(a) Two-person subset: identity diagnostics

Schedule	$\mathrm{sim}_{\mathrm{ref}} \uparrow$	$\mathrm{sim}_{\mathrm{GT}} \uparrow$	Arc-ref ↑	Arc-GT ↑	$\mathrm{sim}_{\mathrm{mis}} \downarrow$
Constant 2.0	0.6207	0.4254	0.6234	0.4211	0.1390
Layer $2 \to 1.75$	0.6153	0.4253	0.6162	0.4184	0.1389
Layer $2 \to 1.5$	0.6016	0.4195	0.5993	0.4082	0.1399
Layer $2 \to 1.0$	0.5748	0.4053	0.5646	0.3880	0.1428
Step $2 \to 1.0$	0.5824	0.4095	0.5723	0.3915	0.1405
Joint layer×step	0.5502	0.3937	–	–	–

(b) Two-person subset: reference preference and global quality

Schedule	$\mathrm{M}_{\mathrm{CP}} \downarrow$	Aes. ↑	CLIP-I ↑	CLIP-T ↑	FID ↓
Constant 2.0	0.2379	5.5010	0.7771	0.3322	116.18
Layer $2 \to 1.75$	0.2330	5.4984	0.7804	0.3317	114.97
Layer $2 \to 1.5$	0.2212	5.5010	0.7783	0.3315	115.45
Layer $2 \to 1.0$	0.2046	5.5276	0.7823	0.3316	115.83
Step $2 \to 1.0$	0.2078	5.5182	0.7786	0.3320	115.74
Joint layer×step	0.1880	–	–	–	115.43

(c) Three-/four-person subset: identity diagnostics

Schedule	$\mathrm{sim}_{\mathrm{ref}} \uparrow$	$\mathrm{sim}_{\mathrm{GT}} \uparrow$	Arc-ref ↑	Arc-GT ↑	$\mathrm{sim}_{\mathrm{mis}} \downarrow$
Constant 2.0	0.6345	0.4090	0.5876	0.3598	0.1067
Layer $2 \to 1.75$	0.6229	0.4087	0.5547	0.3586	0.1077
Layer $2 \to 1.5$	0.6128	0.4033	0.5524	0.3474	0.1062
Layer $2 \to 1.0$	0.5570	0.3881	0.4946	0.3234	0.0987
Step $2 \to 1.0$	0.6047	0.4042	0.5538	0.3488	0.1041
Joint layer×step	0.5436	0.3832	–	–	–

(d) Three-/four-person subset: reference preference and global quality

Schedule	$\mathrm{M}_{\mathrm{CP}} \downarrow$	Aes. ↑	CLIP-I ↑	CLIP-T ↑	FID ↓
Constant 2.0	0.2761	5.3730	0.7613	0.3163	187.39
Layer $2 \to 1.75$	0.2778	5.3589	0.7723	0.3164	186.13
Layer $2 \to 1.5$	0.2757	5.4321	0.7696	0.3159	186.14
Layer $2 \to 1.0$	0.2260	5.4667	0.7650	0.3193	184.94
Step $2 \to 1.0$	0.2626	5.3962	0.7654	0.3182	184.86
Joint layer×step	0.1938	–	–	–	187.70

The two crop controls instead replace each full reference with its resized face crop, either without routing or with all keys of that cropped reference biased.

Relative to the full-reference baseline, face-local routing raises $\mathrm{sim}_{\mathrm{ref}}$ by 0.0646, $\mathrm{sim}_{\mathrm{GT}}$ by 0.0263, ArcFace reference similarity by 0.0733, and ArcFace ground-truth similarity by 0.0327. CLIP-I changes by $+0.0003$, while Aesthetic decreases by 0.0444. These aggregate differences describe the measured trade-off; without paired uncertainty or a non-inferiority margin, they do not establish preservation of global image quality.

Face-local routing does not dominate whole-reference routing on every metric. It is better on the paper's primary reference-identity measures ($+0.0083$ $\mathrm{sim}_{\mathrm{ref}}$ and $+0.0064$ ArcFace-ref), slightly better on $\mathrm{sim}_{\mathrm{GT}}$ ($+0.0034$), and reduces CP by 0.0072. Whole-reference routing is better on ArcFace-GT ($+0.0050$), CLIP-I ($+0.0022$), and CLIP-T ($+0.0005$), while face-local routing is better on Aesthetic ($+0.0413$). This comparison supports a specific claim: restricting the biased *key scope* to faces gives a better identity-retention/reference-preference Pareto point for our primary objective, rather than a universal claim that local processing is superior to global conditioning.

Input cropping is not an equivalent implementation of face routing. The crop controls use the same detector-derived face box and crop construction used by the inference preprocessing; the distinction

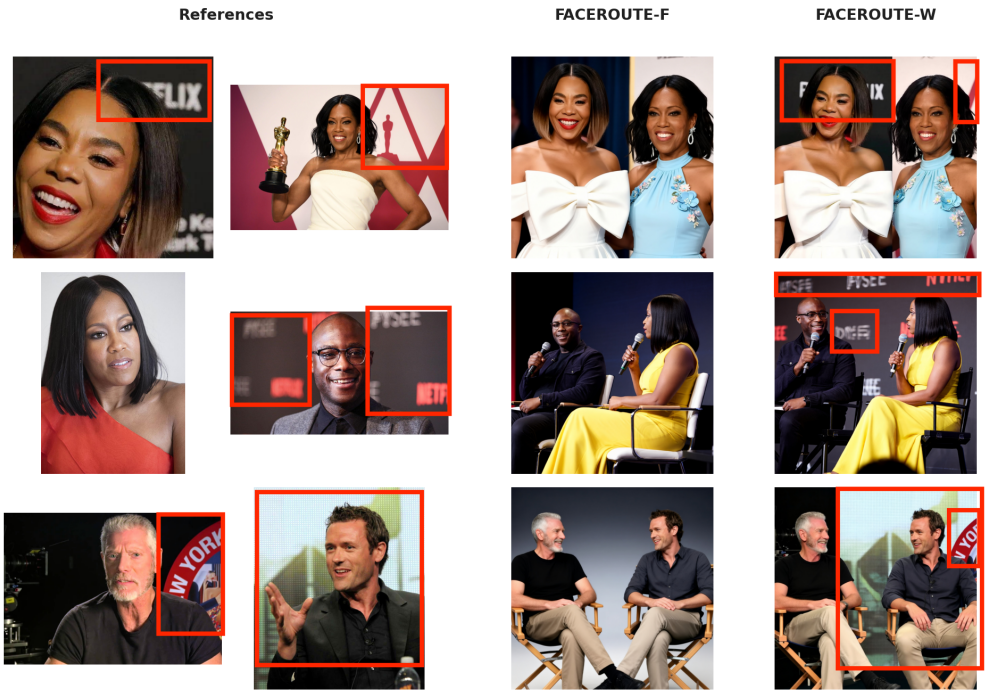

Figure 11: Three selected P2 comparisons on Qwen at $\alpha = 1.5$. Each row shows two full references, FACEROUTE-F, and FACEROUTE-W. Red boxes mark reference-background cues and their appearance in the FACEROUTE-W output; these cues are not required by the target prompts. The examples illustrate the background-transfer ratings, not their frequency.

is that they replace the complete reference image, whereas FACEROUTE-F keeps the complete reference and changes only the attention prior on its face keys. Replacing the natural references with enlarged face crops causes the baseline $\mathrm{sim}_{\mathrm{ref}}$ to fall by 0.3377 and ArcFace-ref by 0.3466; CLIP-I falls by 0.0361 and Aesthetic by 0.4284. Routing the cropped references recovers some identity (e.g., $+0.0468$ $\mathrm{sim}_{\mathrm{ref}}$ over crop-only) but remains far below either full-reference method. A likely explanation is the combination of information removal and conditioning shift: tight crops discard hairstyle extent, body shape, clothing, pose, and scene cues, while resizing a face to the full reference canvas changes its scale and VAE-token statistics relative to the natural images consumed by the editing backbone. The CLIP-I and Aesthetic decreases show that the measured changes extend beyond face similarity, but do not isolate which preprocessing change causes them. The crop rows' low M_{CP} accompanies much lower identity similarity and should not be interpreted as favorable identity preservation.

We repeat this stress test with FLUX.2-KLEIN-BASE-9B using a FLUX conversion of the same aligned 31-example source manifest (Table 20). Relative to the FLUX baseline, face-local routing raises $\mathrm{sim}_{\mathrm{ref}}$ by 0.0255 and $\mathrm{sim}_{\mathrm{GT}}$ by 0.0134. At matched $\alpha = 1.5$, it also exceeds whole-reference routing by 0.0046 in $\mathrm{sim}_{\mathrm{ref}}$ and 0.0075 in $\mathrm{sim}_{\mathrm{GT}}$, while reducing CP by 0.0060 and improving Aesthetic by 0.0847; CLIP-I changes by $+0.0024$ and CLIP-T by -0.0029. These are fixed-seed aggregate differences without paired uncertainty and are therefore descriptive, not evidence of statistical significance.

On the full 49-example Qwen small-face stress set, FACEROUTE-F exceeds FACEROUTE-W by $+0.0086$ in $\mathrm{sim}_{\mathrm{ref}}$ and $+0.0024$ in $\mathrm{sim}_{\mathrm{GT}}$ under the pooled, metric-wise IQR aggregation used in Figure 3. With detected faces occupying 1.63% of the reference canvas on average, these results support face-local reweighting in the tested small-face setting. The contrast with ordinary Qwen P2 is consistent with concentrating the added prior on scarce face tokens; the two datasets also differ in content.

Official key-scope control. The local stress test does not establish that face-local keys are the best scope on the official benchmark or on another backbone. We therefore hold the backbone, split, target-query scope, strength ($\alpha = 1.5$), seed, and valid IDs fixed and change only the biased reference-key set. Table 21 reports the official MultiID-Bench aggregates; Appendix C gives exact-common-ID paired uncertainty.

The scope effect is not universal. On Qwen p2, whole-reference routing gives a resolved paired gain in $\mathrm{sim}_{\mathrm{ref}}$ ($+0.0113$, 95% CI $[+0.0031, +0.0202]$) and a resolved higher CP cost ($+0.0115$, $[+0.0036, +0.0196]$). Its $\mathrm{sim}_{\mathrm{GT}}$ gain ($+0.0043$, $[-0.00004, +0.0087]$) is unresolved. The aggregate FID is 0.52 lower and CLIP-I is 0.0036 higher for whole-reference routing, while Aesthetic and CLIP-T are lower; these non-paired quality differences are descriptive. On FLUX p2, whole-reference routing gives small but resolved paired gains over face-local routing in $\mathrm{sim}_{\mathrm{ref}}$ ($+0.0077$, 95% CI $[+0.0029, +0.0126]$) and $\mathrm{sim}_{\mathrm{GT}}$ ($+0.0074$, $[+0.0035, +0.0114]$); its CP difference ($+0.0028$, $[-0.0016, +0.0074]$) is unresolved. On FLUX p3, all three paired intervals cross zero. For Qwen p3, the configured benchmark aggregate favors face-local routing in both identity measures, but the untrimmed paired identity differences are unresolved. In contrast, whole-reference routing has a resolved higher CP cost ($+0.0283$, $[+0.0093, +0.0467]$; 22 CP-valid pairs). Accordingly, the cross-backbone evidence supports a general *reference-access* effect. Face localization acts as a selective prior that can reduce the identity–reference-preference cost of broader reweighting, but it is not a universally optimal or uniquely causal key scope.

M.7 Detector-Box Margin Sensitivity

The default method maps the unmodified SCRFD box (scale 1.0) to packed VAE tokens. To test sensitivity to the box margin, we isotropically scale every detected box about its center by 0.8 or 1.2, clip the result to the VAE canvas, and regenerate the complete Qwen two-person split. All other settings match the default 40-step, all-target, face-local configuration with $\alpha = 1.5$ and the same seed. We compare exact common sample IDs without IQR trimming and obtain 95% confidence intervals from $10{,}000$ paired bootstrap replicates.

Table 24: Sensitivity to the detector-box margin on Qwen MultiID-Bench p2. Each entry is the paired change from the default box scale 1.0; brackets give 95% bootstrap confidence intervals. Higher is better for identity and AES, whereas lower is better for M_{CP}.

Box scale	n	$\Delta\mathrm{sim}_{\mathrm{ref}}$	$\Delta\mathrm{sim}_{\mathrm{GT}}$	$\Delta\mathrm{M}_{\mathrm{CP}}$	ΔAES
0.8	194	-0.0006 $[-0.0053, 0.0050]$	-0.0012 $[-0.0052, 0.0030]$	$+0.0028$ $[-0.0017, 0.0073]$	-0.0077 $[-0.0294, 0.0141]$
1.2	193	$+0.0038$ $[-0.0019, 0.0094]$	-0.0006 $[-0.0044, 0.0032]$	$+0.0075$ $[+0.0017, +0.0135]$	-0.0040 $[-0.0234, 0.0158]$

Relative to the unrouted backbone, scales 0.8, 1.0, and 1.2 improve $\mathrm{sim}_{\mathrm{ref}}$ by $+0.0446$, $+0.0451$, and $+0.0494$, respectively, and improve $\mathrm{sim}_{\mathrm{GT}}$ by $+0.0250$, $+0.0262$, and $+0.0259$. The identity intervals in Table 24 all cross zero, as do the AES, CLIP-I, and CLIP-T contrasts (the latter two omitted for compactness). Expanding the box to 1.2 slightly increases M_{CP} relative to scale 1.0; this is a small preference-cost change without a resolved additional $\mathrm{sim}_{\mathrm{GT}}$ gain. Overall, the primary identity improvement is stable within the tested $\pm 20\%$ margin range. Because all variants start from the same SCRFD detections, this experiment measures box-margin sensitivity rather than robustness to detector choice, false detections, or missed faces.

M.8 Reference Isolation

Figure 9(a) and Table 22 compare reference-isolation variants on the 40-step p3 setting with $\mathcal{T}_{\mathrm{all}}$ and $\alpha = 2$ fixed. Here "none" means *routing without isolation*, not the baseline. Full isolation reduces $\mathrm{sim}_{\mathrm{ref}}$ from 0.6345 to 0.5164. Restricting isolation to face tokens recovers part of the loss, reaching 0.5770, but still underperforms no isolation and worsens $\mathrm{sim}_{\mathrm{mis}}$. Lower M_{CP} under isolation is therefore obtained by suppressing useful identity transfer rather than by improving the overall trade-off.

M.9 Layer and Step Scheduling

Single-axis schedules soften the identity/copy trade-off, but their preferred endpoint varies by split and inference stack. Under the 40-step Qwen base configuration with $\alpha = 2$, p2 layer $2 \to 1.75$ retains most of the constant identity score with a lower observed FID. On p3, step $2 \to 1$ has the lowest observed FID among the listed single-axis schedules while retaining a substantial identity gain. These within-table FID rankings are descriptive. By contrast, the multiplicative joint layer–step schedule nearly returns to baseline on both splits ($\mathrm{sim}_{\mathrm{ref}}$ 0.5502 and 0.5436). The two simultaneous decays make the effective bias weak over most layer/step pairs.

M.10 Where Does Routing Matter?

Figure 9(b) shows that, on the 40-step p3 setting at $\alpha = 2$, early-only, middle-only, and late-only routing obtain $\mathrm{sim}_{\mathrm{ref}}$ scores of 0.5400, 0.5721, and 0.6163, respectively. This ordering suggests that later transformer blocks carry most of the measured identity benefit from late identity routing. This analysis uses the same 31 valid p3 examples and should be read as a mechanistic observation rather than a universal claim about transformer depth.

M.11 Qualitative Comparison

Figure 1 shows three selected p2 examples at matched seeds. They compare the unmodified editor with whole-reference and face-local routing at $\alpha = 1.5$. Face-local routing retains full-reference access while increasing the attention prior only on face keys; it does not prevent target queries from reading other parts of the reference image. The examples are selected for visual clarity only; all quantitative comparisons use the complete aligned p2 benchmark.

Figure 10 shows two selected examples from the 49-example Qwen small-face stress test. In both cases, face-local routing improves visible facial identity rendering relative to the unmodified baseline. These examples are illustrative only; quantitative scope comparisons use all 49 examples (Figure 3).

N Geometry-Based Reference-Preference Diagnostic

For each assigned generated, ground-truth, and reference face triplet (g, t, r), we compute 106-point landmark distances after similarity-Procrustes alignment and distances between head-pose vectors $\boldsymbol{\theta} = (\theta_{\mathrm{pitch}}, \theta_{\mathrm{yaw}}, \theta_{\mathrm{roll}})^\top$. For either distance d, reference preference is

$$\rho_d = \frac{d(g,t) - d(g,r)}{d(g,t) + d(g,r)}. \tag{9}$$

Positive values indicate geometry closer to the reference than to the ground truth. This replaces the recognition-similarity descriptor, not the learned detection and face-assignment stages. Intervals spanning zero indicate an unresolved difference, not equivalence or proof that copying is absent.

Table 25 reports the paired landmark and pose reference-preference changes versus the unrouted 40-step Qwen baseline.

O Identity-Blending and Face-Collapse Audit

This appendix reports identity separability (the margin between the assigned and the closest wrong reference), face collapse, and image-level ambiguity, together with per-sample worst cases.

Metrics. For each generated face we form a set-level one-to-one (Hungarian) match between generated and reference faces. Let s_i^+ be the similarity to the matched identity and $s_i^- = \max_{j \neq i} \mathrm{sim}(g_i, r_j)$ the largest wrong-reference similarity for face i. At image level, we report (i) the mean *identity margin* $\bar{m} = K^{-1} \sum_i (s_i^+ - s_i^-)$ (higher is better); (ii) the worst competing-identity similarity $s_{\max}^- = \max_i s_i^-$ (lower is better); and (iii) a *collapse excess* $e = \max_{i \neq j} \mathrm{sim}(g_i, g_j) - c^{\mathrm{ref}}$, where c^{ref} is the analogous maximum among the reference faces, providing a reference-similarity offset rather than a binary collapse test; and (iv) an *image-level ambiguity rate*. We first calibrate a

Table 25: Landmark and pose reference-preference changes versus the unrouted 40-step Qwen baseline. Higher positive values indicate more reference-like geometry; brackets are unadjusted paired-bootstrap 95% CIs. Although three intervals at $\alpha = 2.5$ exclude zero before correction, none of the corresponding tests survive Holm correction.

Split	α	n	Δ landmark pref.	Δ pose pref.
p2	1.5	187	$+0.0121$ $[-0.0036, +0.0290]$	$+0.0136$ $[-0.0018, +0.0318]$
p2	2.5	186	$+0.0238$ $[+0.0015, +0.0458]$	$+0.0190$ $[-0.0020, +0.0398]$
p3	1.5	26	$+0.0206$ $[-0.0229, +0.0696]$	$+0.0103$ $[-0.0301, +0.0568]$
p3	2.5	27	$+0.0705$ $[+0.0119, +0.1335]$	$+0.0603$ $[+0.0067, +0.1180]$

per-panel threshold as the 10th percentile of all baseline face margins, mark an image as ambiguous if at least one generated face falls below that threshold, and report the fraction of ambiguous images. The threshold is calibrated on the retained baseline faces and held fixed during bootstrap resampling, so these intervals do not include calibration uncertainty. Ambiguity is a data-adaptive diagnostic, not an externally calibrated failure probability. All quantities use ArcFace on strict samples with equal, nonzero generated and reference face counts; baseline and $\alpha = 1.5$ are paired on exact common IDs. Reference faces are reindexed according to the Hungarian assignment in each arm; this set-level assignment is not an identity-to-position check. For a given sample, c^{ref} is shared between arms and cancels in the paired difference. Hence Δe equals the change in raw maximum generated-face similarity, not an independent collapse result.

Retained and excluded samples. The strict intersection retains 182 of 194 Qwen two-person samples, 20 of 31 Qwen three-/four-person samples, 180 of 194 FLUX two-person samples, and 25 of 31 FLUX three-/four-person samples. Thus 12, 11, 14, and 6 samples, respectively, are excluded because at least one arm fails the face-count criterion. A mismatch can reflect generation or detection failure; this audit does not distinguish the two or estimate the effect on missing-face frequency. Results below are conditional on the common retained subset.

Table 26: Identity-blending audit for FACEROUTE-F with all-target queries at $\alpha = 1.5$ versus the unrouted baseline (ArcFace, sample-paired on exact common IDs). Higher mean identity margin and lower worst competing-identity similarity, collapse excess, and ambiguity are better. Intervals are paired-bootstrap 95% CIs (10,000 replicates; seed 20260901). Ambiguity is reported as a percentage-point change in the image-level rate.

Backbone	Split	n	$\Delta\bar{m}$ [95% CI]	$\Delta s^{-}_{\max}$ [95% CI]
Qwen	2-person	182	$+0.0598$ $[+0.0495, +0.0707]$	-0.0031 $[-0.0115, +0.0050]$
Qwen	3–4-person	20	$+0.0681$ $[+0.0461, +0.0886]$	$+0.0196$ $[-0.0037, +0.0503]$
FLUX	2-person	180	$+0.0551$ $[+0.0474, +0.0633]$	-0.0015 $[-0.0067, +0.0036]$
FLUX	3–4-person	25	$+0.0893$ $[+0.0753, +0.1033]$	-0.0037 $[-0.0155, +0.0083]$

Backbone	Split	n	Δe [95% CI]	Δ ambiguity (pp) [95% CI]
Qwen	2-person	182	-0.0309 $[-0.0406, -0.0213]$	-4.95 $[-9.89, -0.55]$
Qwen	3–4-person	20	-0.0059 $[-0.0419, +0.0351]$	-5.00 $[-25.00, +15.00]$
FLUX	2-person	180	-0.0203 $[-0.0284, -0.0122]$	-7.22 $[-11.67, -2.78]$
FLUX	3–4-person	25	-0.0347 $[-0.0523, -0.0161]$	-12.00 $[-24.00, +0.00]$

Result. Across all four panels the mean identity margin rises with paired intervals excluding zero. The change in worst competing-identity similarity is unresolved in every panel: intervals include both zero and positive values. This is not an equivalence or non-inferiority result and does not rule out a meaningful increase, particularly in the Qwen three-/four-person panel, whose upper limit is $+0.0503$. Collapse excess decreases on three of four panels and is inconclusive only on the smallest (Qwen three-/four-person, $n = 20$). Image-level ambiguity drops on both two-person panels. Table 26 summarizes the deltas. In words, FACEROUTE-F increases average separation between assigned and competing identities on the retained subset. Increased margins alone do not certify that a face is perceptually unmixed: assigned-reference similarity can rise even when competing similarity also rises. These single-seed ArcFace results therefore neither establish population-level immunity to blending nor replace human judgments of hybrid facial appearance.

Worst cases. Qwen three-/four-person sample `9fa7bb8f` degrades on all three diagnostics ($\Delta(\min_i m_i) = -0.288$, $\Delta s^{-}_{\max} = +0.249$, and $\Delta e = +0.248$). Collapse also rises in individual two-person cases (e.g., Qwen `47e931c2`, $\Delta e = +0.156$; FLUX `cb8be432`, $\Delta e = +0.224$). These failures coexist with the aggregate margin gain; face localization is not uniformly protective.

P XVERSEBENCH PROTOCOLS AND RETENTION–QUALITY TRADE-OFFS

Same-backbone setup. We evaluate all 300 XVerseBench examples (90 single-, 120 two-, and 90 three-subject) using non-distilled FLUX.2-KLEIN-BASE-9B, no LoRA, Diffusers 0.37.0, bfloat16, and CPU model offload. The routed arms use FACEROUTE with SAM3 subject-mask key selection. Baseline and all five routed arms use 1024×1024 outputs, 40 steps, CFG 4.0, and seed 42 reset per prompt. Official prompts and reference order are unchanged. Full RGB references undergo standard resizing and VAE-grid alignment, without subject cropping or background removal.

Mask-to-key mapping. All routed arms reuse 134 SAM3 masks (45 animal, 20 human, 69 object), with three accessory corrections. We select a token when foreground covers at least half of its 16×16 pixel footprint on the aligned reference grid. The additive logit bias is $B_{qk}(s) = b(s)\mathbb{1}[q \in$

$\mathcal{T}, k \in \mathcal{F}$], applied at all 32 layers and both CFG branches. Fixed-strength arms use $b(s) = \log \alpha$, $\alpha \in \{1.5, 2.0, 2.5\}$, at every step. No negative bias or token-count normalization is used.

Time-varying bias. With one-based inference step $s = 1, \ldots, 40$ (not the scheduler timestep), the two exploratory schedules are

$$b_{\mathrm{P}}(s) = \begin{cases} 1 - \dfrac{s-10}{15} + \dfrac{14}{135}(s-10)^2, & 1 \le s \le 10, \\ \left(\dfrac{40-s}{30}\right)^2, & 10 < s \le 40, \end{cases} \tag{10}$$

$$b_{\mathrm{Q}}(s) = 5\left(\frac{40-s}{39}\right)^2. \tag{11}$$

Piecewise decay has $b_{\mathrm{P}}(1) = 10$, $b_{\mathrm{P}}(10) = 1$, and $b_{\mathrm{P}}(40) = 0$; quadratic decay starts at 5 and ends at 0. These values are logit biases, not multipliers: the initial attention-odds multipliers are e^{10} and e^5. Although quadratic decay starts lower, it exceeds the piecewise bias at some later steps. The schedules do not match integrated bias or attention mass, so their comparison cannot separate strength from timing.

Cross-model reproductions. XVerse (Chen et al., 2025) uses FLUX.1-dev (Black Forest Labs, 2024) with its released adapters, 1024^2 outputs, 28 steps, and guidance 3.5. Original RGB sources pass through native square padding and the CLIP 224-pixel and VAE 256-pixel reference branches. Both MOSAIC (She et al., 2026) runs use FLUX.1-dev and the same released `subject_512` LoRA at scale 1.0. The 512^2 run uses 28 steps, guidance 3.5, and SAM3 foregrounds composited on white, cropped, and padded. The 1024^2 run uses 40 steps, guidance 4, and original RGB sources through MOSAIC's native preprocessing. Reference size equals output size and reference scale is 0.9 in both. Thus this pair is not an isolated resolution ablation, and the second run does not use a separately trained 1024-pixel adapter. All nine configurations share the same prompts, ordered reference identities, and numerical seed; the latter does not imply identical initial noise across models.

Evaluation protocol. Released XVerse inference uses segmented references and four generated images per prompt, saved as a 2×2 grid and split for evaluation. Our shared adapter uses original references and one output per prompt with the released metric algorithms. Scores are not directly comparable to published XVerse results.

Measures and denominators. For non-human references, IP uses Florence-2 detection and SAM2 segmentation, retaining up to three generated candidates. Square-padded crops are resized to 256 and center-cropped to 224 for DINO ViT-S/16 cosine similarity. Human ID uses RetinaFace-ResNet50, $1.3\times$ box expansion, the released alignment, and IR-SE-50 cosine similarity. Each reference takes its highest candidate similarity, scaled by 100, without one-to-one assignment. DPG uses mPLUG VQA with the released dependency rule on the original task questions. AES uses Aesthetic Predictor V2.5 with a SigLIP encoder in bfloat16, multiplying its output by 10; it is a learned proxy, not a human rating or accuracy. IP and ID average 400 non-human and 200 human reference occurrences, respectively; DPG and AES average 300 images. The single-, two-, and three-subject IP counts are 70, 160, and 170; ID counts are 20, 80, and 100.

Detector accounting. Missing detections remain as zero scores, keeping identical denominators. In the order baseline, $\alpha = 1.5$, 2.0, 2.5, piecewise decay, quadratic decay, XVerse, MOSAIC 512^2, and MOSAIC 1024^2 (Tables 27 and 28), IP no-detection counts are $(4, 4, 3, 5, 2, 3, 0, 0, 3)$ and ID counts are $(2, 2, 2, 2, 5, 4, 3, 3, 4)$. These count zero detected candidates, not all possible alignment failures. Evaluation always uses original RGB references, including for the masked-input MOSAIC run. Generation SAM3 masks and evaluation SAM2 masks are obtained independently. Table 29 breaks the same-backbone scores out by subject count.

Cross-model comparison. Table 27 lists the local XVerse and MOSAIC reproductions on the same 300 prompts and evaluator. Their FLUX.1-dev backbones, learned adapters, and generation settings differ from our controlled arms, and the two MOSAIC configurations also differ in reference preprocessing and sampling, not just resolution. These rows are therefore neither a routing ablation nor published leaderboard results.

Table 27: XVerseBench local reproductions and same-backbone FACEROUTE (mask) comparisons, 300 prompts per row. All metrics are higher-is-better; IP/ID average 400/200 reference occurrences and DPG/AES average 300 images, with missing detections scored as zero. FACEROUTE (mask) uses SAM3 subject-mask keys. Cross-system backbones, sampling, and preprocessing differ; these are not published leaderboard scores.

Method	IP ↑	ID ↑	DPG ↑	AES ↑
FLUX.1-dev with learned personalization modules				
XVerse, 1024^2	62.40	63.38	92.21	62.47
MOSAIC, 512^2, masked refs.	63.61	38.51	92.05	59.65
MOSAIC, 1024^2, original refs.	56.26	34.99	88.99	60.11
FLUX.2-klein-base-9B without extra training, 1024^2				
Unmodified backbone	66.59	50.12	94.42	62.92
FACEROUTE (mask), $\alpha = 1.5$	68.11	54.68	94.93	62.44
FACEROUTE (mask), $\alpha = 2.5$	69.65	60.09	93.93	61.76

Additional routing settings. Table 28 completes the same-backbone sweep. Exploratory decay schedules produce larger ID gains but lower DPG and AES; quadratic decay barely changes IP. Because both strength and timing change, these runs do not isolate a timing advantage.

Table 28: Additional same-backbone FACEROUTE (mask) settings on all 300 XVerseBench prompts, using the FLUX.2-klein-base-9B protocol and evaluator of Table 27, which also reports Baseline and $\alpha \in \{1.5, 2.5\}$. All metrics are higher-is-better.

Method	IP ↑	ID ↑	DPG ↑	AES ↑
FACEROUTE (mask), $\alpha = 2.0$	69.52	58.10	94.28	61.92
FACEROUTE (mask), piecewise decay	69.02	67.64	90.88	58.44
FACEROUTE (mask), quadratic decay	66.78	76.65	88.57	56.24

Table 29: Same-backbone XVerseBench FACEROUTE (mask) scores by subject count, including both dynamic schedules. All metrics are higher-is-better; parentheses count images. IP/ID denominators differ from image counts as specified above. Full-set scores are in Tables 27 and 28.

Group	Method	IP ↑	ID ↑	DPG ↑	AES ↑
Single (90)	Unmodified backbone	69.96	63.07	97.50	63.43
	FACEROUTE (mask), $\alpha = 1.5$	71.78	64.58	97.22	62.89
	FACEROUTE (mask), $\alpha = 2.0$	72.71	69.40	97.50	62.34
	FACEROUTE (mask), $\alpha = 2.5$	72.98	70.18	97.22	62.42
	FACEROUTE (mask), piecewise decay	78.73	82.45	94.07	60.61
	FACEROUTE (mask), quadratic decay	80.74	88.56	93.80	60.61
Two (120)	Unmodified backbone	69.28	53.70	93.60	62.84
	FACEROUTE (mask), $\alpha = 1.5$	69.85	58.80	93.89	62.19
	FACEROUTE (mask), $\alpha = 2.0$	71.09	62.70	93.58	61.76
	FACEROUTE (mask), $\alpha = 2.5$	71.54	64.56	92.93	61.66
	FACEROUTE (mask), piecewise decay	68.41	68.56	90.40	58.12
	FACEROUTE (mask), quadratic decay	65.22	81.38	89.03	55.42
Three (90)	Unmodified backbone	62.68	44.67	92.42	62.51
	FACEROUTE (mask), $\alpha = 1.5$	64.97	49.41	94.04	62.33
	FACEROUTE (mask), $\alpha = 2.0$	66.74	52.16	92.01	61.70
	FACEROUTE (mask), $\alpha = 2.5$	66.49	54.49	91.98	61.23
	FACEROUTE (mask), piecewise decay	65.58	63.93	88.33	56.70
	FACEROUTE (mask), quadratic decay	62.49	70.48	82.74	52.96

Subject-count dependence. Every fixed-strength arm improves IP and ID over baseline in each subject-count group; this does not hold for dynamic routing. Piecewise decay reduces two-subject IP from 69.28 to 68.41, and quadratic decay reduces it to 65.22, despite larger ID gains. Quadratic

decay also lowers three-subject DPG from 92.42 to 82.74 and AES from 62.51 to 52.96. These group differences caution against judging a schedule by aggregate ID alone. The study is single-seed, without held-out schedule selection or significance testing; $\alpha = 1.5$ remains the default.

Null check. One single-subject smoke test at $\alpha = 1$ matches the native baseline in pixels and PNG checksum; this is not a full-set audit.

Interpretation limits. Best-candidate similarity does not establish separation or accessory fidelity: an inspected output omitted an accessory included in its mask. Without a copy-paste diagnostic or blinded human evaluation, these gains do not establish fine-detail preservation or absence of copying.

Q Exploratory Quality-Guidance Smoke Tests

Scope and control. We test whether quality-oriented text conditioning can improve AES without lowering subject similarity or original-task DPG. Three successive, adaptively designed rounds reuse the same previously inspected XVerseBench prompts: a polar bear (`xverse_single_00000`), a man riding a motorcycle (`xverse_multi_00000`), and a man, house, and wolf (`xverse_multi_00120`). Control A is FaceRoute with subject-mask routing at $\alpha = 1.5$, not the unmodified backbone. All arms retain original RGB references, SAM3 masks, all-target routing at every layer and step in both CFG branches, 40 steps, CFG 4, seed 42, and 1024^2 outputs. Control images and main metrics are reused after hash checks, not independently regenerated. The nine treatments add 27 images, not 27 independent prompts; these tests are hypothesis-generating only.

Quality text and negative CFG. B appends the following suffix to the original positive prompt:

> Coherent natural lighting across the scene, consistent light direction, realistic contact shadows, physically plausible relative scale, and seamless integration of the subjects with their surroundings.

C retains the original positive prompt and replaces the empty negative text with:

> Inconsistent lighting, conflicting light directions, missing contact shadows, implausible relative scale, pasted-on subjects, visible cutout edges, disconnected foreground and background.

D combines B and C. Both CFG branches retain the same reference images and routing. No explicit pose or camera-framing intervention is tested.

Light-only conditioning and timing. E–J use empty negative text and the shorter positive suffix:

> Coherent natural lighting with a consistent light direction and realistic contact shadows.

Let p_0 and p_L be the padded conditional embeddings of the original and light-suffixed prompts. E uses p_L at all steps; F uses $\frac{1}{2}p_0 + \frac{1}{2}p_L$ throughout. With zero-based step $s = 0, \ldots, 39$, G, H, and I switch from p_0 to p_L at steps 20, 25, and 30, respectively. J uses p_0 through step 19, then $(1 - w_s)p_0 + w_s p_L$, where $w_s = [1 - \cos(\pi(s - 20)/19)]/2$. Embedding interpolation is performed in bfloat16 and is not assumed linear in image content or model velocity. Runtime audits check the positive and negative embeddings at every CFG call; DPG continues to evaluate the original task questions.

Joint screen. Each round fixes its treatments and decision rule before its new generations. A candidate passes on an image only if every applicable per-reference ID/IP score and DPG are at least A's, and AES strictly increases under both the main bfloat16 evaluator and a float32 rescore of the same predictor. No improvement in one subject compensates for a regression in another. This deterministic screen is not a statistical non-inferiority test. Float32 is a precision diagnostic, not an independent aesthetic measure or a human judgment. Table 30 lists every treatment.

Table 30: All quality-guidance treatments on the same three selected prompts. Means use four non-human references (IP), two human references (ID), and three images (DPG/AES). AES here is bfloat16; the joint count additionally requires float32 AES improvement and per-reference retention. Means do not determine joint success. A is the reused routed control.

Arm	Treatment	IP	ID	DPG	AES	Joint
A	Original prompt	65.95	46.95	100.00	68.65	—
B	Broad positive suffix	60.21	60.44	86.67	70.62	0/3
C	Negative text only	61.55	36.82	100.00	66.56	0/3
D	Positive + negative	62.73	27.00	100.00	68.12	0/3
E	Light suffix, all steps	59.75	43.85	86.67	71.35	0/3
F	Half embedding blend	68.75	34.63	100.00	70.00	0/3
G	Light switch at step 20	66.26	48.77	100.00	69.48	1/3
H	Light switch at step 25	66.45	48.76	100.00	68.75	1/3
I	Light switch at step 30	66.09	48.65	100.00	68.85	1/3
J	Cosine ramp, steps 20–39	66.61	49.06	100.00	69.58	1/3

Findings and limits. B–F produce no joint successes; C lowers AES on all three images in both precisions. G–I pass only on the man–house–wolf prompt, while J passes only on the polar bear. For G on the three-subject prompt, ID rises by 3.98, house/wolf IP by $1.76/0.22$, and AES by 3.125 (bfloat16) and 2.41 (float32), with DPG unchanged. However, no treatment passes on more than one prompt. H–J also fail the prespecified follow-up criterion of passing at least two prompts or weakly dominating G on all per-image subject scores, DPG, and both AES precisions with a strict AES gain. The default remains A.

These selected, single-seed observations do not establish general AES improvement, failure of negative prompting in general, or preservation of pose and composition. DPG is already 100 for A on every prompt, and does not measure compliance with the added quality prose. Reusing the same prompts across adaptive rounds precludes treating later rounds as independent validation. The bundled wording does not causally isolate lighting, scale, or integration as sources of low AES; blinded human evaluation and fresh, stratified multi-seed tests remain necessary.

R ALL-INPUT COMPLETENESS AUDIT

The identity-blending audit of Appendix O conditions on samples whose generated and reference face counts match in both arms, so it cannot count subjects that are missing outright. This appendix scores every requested subject in predefined eligible paired cohorts, without generated-face-count filtering. It tests whether the direction of the identity-retention benefit persists under an unconditional, detector-based outcome; it does not prove that every conditional metric is free of selection bias.

Protocol. For each sample we detect faces in the generated image (InsightFace buffalo_l, detection threshold 0.45, detection size 640) and embed detected and reference faces with ArcFace. Both inputs are center-padded to a black square of side $1.5\max(H, W)$ before detection. A decision threshold τ is calibrated per panel from within-sample cross-identity reference impostor pairs at a target false-positive rate (primary $\mathrm{FPR} = 1\%$; ties excluded conservatively). Detected faces are assigned to subjects by Hungarian matching on the cosine matrix. A subject is *covered* when its assigned face reaches τ; otherwise the failure is typed as *confusion* (some detected face reaches τ for this subject but the one-to-one assignment does not), *missing face* (no face left to assign), or *low similarity* (an assigned face exists but no detected face reaches τ). We report *subject recall* within each image, averaged equally across images, and the *image-complete* rate (all requested subjects covered), paired on exact common IDs between the unrouted baseline and face routing at $\alpha = 1.5$ (all-target rows). Brackets are paired-bootstrap 95% CIs over images ($10{,}000$ replicates; seed 20260818), retaining all subject slots in each resampled image. Intervals are pointwise, untrimmed, and conditional on the calibrated threshold; calibration uncertainty is not resampled. The unconditional identity score is $U_i = K_i^{-1} \sum_{k=1}^{K_i} \max(0, s_{ik}^{\mathrm{assigned}})$, with unassigned slots zero, followed by an equal-weight image average. Unlike a per-subject best match, this score respects one-to-one assignment.

Cohorts. Each two-person panel contains 194 images and 388 subject slots. Each three–four-person panel contains 33 images (19 three-person and 14 four-person; 113 slots). This audit needs no ground-truth image and therefore retains the two samples excluded from the original evaluator for missing staged `ori.jpg`; alternate-format ground truths exist in the source benchmark (Appendix B). Small-face uses the fixed 31-ID stress-test cohort for both backbones, excluding 0021, 0030, and prompt-invalid 0032, and recalibrates τ on these references. No reference-extraction or inference failures occurred within these selected cohorts. The older 32-image FLUX audit is retained as a legacy artifact, not mixed into the table.

Table 31: All-input completeness audit at $\alpha = 1.5$ versus baseline (primary $\mathrm{FPR} = 1\%$). Δrecall is the mean paired change in within-image subject recall, with equal image weights; complete counts are images with every requested subject covered. All predefined eligible IDs are retained without generated-face-count filtering.

Backbone	Split	n	Δrecall [95% CI]	Complete (B→R)	Δcomplete [95% CI]
Qwen	2-person (S0)	194	$+0.052$ $[+0.023, +0.080]$	$136 \to 152$	$+0.082$ $[+0.036, +0.129]$
Qwen	2-person (S1)	194	$+0.057$ $[+0.031, +0.082]$	$138 \to 156$	$+0.093$ $[+0.046, +0.139]$
Qwen	2-person (S2)	194	$+0.054$ $[+0.028, +0.082]$	$137 \to 153$	$+0.082$ $[+0.036, +0.129]$
Qwen	3–4-person	33	$+0.048$ $[+0.005, +0.093]$	$14 \to 17$	$+0.091$ $[-0.030, +0.242]$
FLUX	2-person	194	$+0.080$ $[+0.049, +0.111]$	$112 \to 139$	$+0.139$ $[+0.088, +0.191]$
FLUX	3–4-person	33	$+0.048$ $[+0.008, +0.088]$	$13 \to 15$	$+0.061$ $[+0.000, +0.152]$
Qwen	small-face (face)	31	$+0.016$ $[-0.032, +0.081]$	$24 \to 25$	$+0.032$ $[-0.065, +0.161]$
Qwen	small-face (whole)	31	$+0.048$ $[+0.000, +0.097]$	$24 \to 27$	$+0.097$ $[+0.000, +0.194]$
FLUX	small-face (face)	31	$+0.129$ $[+0.065, +0.210]$	$2 \to 5$	$+0.097$ $[+0.000, +0.226]$
FLUX	small-face (whole)	31	$+0.145$ $[+0.048, +0.258]$	$2 \to 6$	$+0.129$ $[+0.032, +0.258]$

Results. Recall and completeness have positive point estimates in all ten panels (Table 31); not all intervals exclude zero. Qwen two-person recall increases by 0.052–0.057 across the three seeds. Respectively, $19/21/19$ images become complete and 3 regress in each seed, giving net gains of $16/18/16$. On FLUX two-person, 28 become complete and 1 regresses. On three–four-person, the completeness interval crosses zero for Qwen and includes zero for FLUX. Qwen's small-face recall and completeness effects also remain unresolved; FLUX small-face recall increases for both scopes, but absolute completeness remains low ($5/31$ and $6/31$ versus $2/31$).

Table 32: Unconditional assigned-identity score and paired transitions. New/lost complete counts are images; new/lost coverage counts are subject slots. Positive net effects do not exclude individual regressions.

Backbone	Split	ΔU [95% CI]	New/lost complete	New/lost coverage
Qwen	2-person (S0)	$+0.057$ $[+0.050, +0.065]$	19/3	26/6
Qwen	2-person (S1)	$+0.053$ $[+0.045, +0.061]$	21/3	27/5
Qwen	2-person (S2)	$+0.063$ $[+0.056, +0.072]$	19/3	25/4
Qwen	3–4-person	$+0.063$ $[+0.037, +0.088]$	4/1	7/2
FLUX	2-person	$+0.054$ $[+0.048, +0.061]$	28/1	35/4
FLUX	3–4-person	$+0.076$ $[+0.063, +0.089]$	2/0	8/2
Qwen	small-face (face)	$+0.075$ $[+0.050, +0.102]$	2/1	3/2
Qwen	small-face (whole)	$+0.071$ $[+0.044, +0.102]$	3/0	3/0
FLUX	small-face (face)	$+0.031$ $[+0.019, +0.047]$	3/0	8/0
FLUX	small-face (whole)	$+0.027$ $[+0.011, +0.045]$	4/0	10/1

The assigned-identity change is positive with intervals above zero in all ten panels (Table 32), supporting improved identity retention without face-count conditioning on these cohorts. Individual coverage regressions remain. At the primary threshold, Qwen three–four-person missing-face failures increase from 3 to 6 of 113 slots, while low-similarity failures fall from 21 to 14 and confusion from 4 to 3. A separate flag—at least two detected faces exceeding τ for one reference—occurs in 15 baseline and 16 routed images there; it is not a verified duplicate identity. Detector misses and actual omissions cannot be distinguished by this audit.

Sensitivity and provenance. On the six main two-/three–four-person panels, recall and completeness point estimates remain positive at $\mathrm{FPR} = 0.5\%$ and 5%. Detection thresholds 0.35 and 0.55 were tested only on Qwen two-person S0 (Δrecall $+0.052$, $+0.049$; complete $136 \to 152$, $136 \to 151$). Pooled subject-weighted recall gives $5/113 = +0.044$ for Qwen three–four-person and $6/113 = +0.053$ for FLUX, versus the image-weighted $+0.048$ in both. Small-face has only 31 reference impostor pairs, so 0.5% and 1% permit no passing pair and yield the same threshold; they are not independent stability checks. The v2 statistics script reconstructs U from assignments and records input hashes, sorted IDs, aggregation rules, and bootstrap settings. It supersedes an earlier summary that mixed best-match and assigned scores; the legacy file is preserved.

Backend-pipeline sensitivity. FaceNet and AdaFace each fail to extract a reference face in two of the 194 Qwen S0 samples. We therefore report the same 192-image intersection separately, not as a complete benchmark replication (Table 33). FaceNet's initial run also had three GPU-memory failures; a retry resolves those infrastructure failures but not the two reference-extraction failures. Both alternative pipelines use RetinaFace/FaceRestoreHelper detection and alignment, so this check changes more than the embedding network. Their absolute scores and thresholds are not comparable to ArcFace's. On this subset, both pipelines retain positive recall, completeness, and U changes, with pointwise intervals above zero.

Table 33: Backend-pipeline sensitivity on Qwen two-person S0, $n = 192/194$, $\alpha = 1.5$. Recall and completeness changes are percentage points (pp); U remains on its original scale. Each pipeline calibrates its own threshold at $\mathrm{FPR} = 1\%$.

Pipeline	Δrecall (pp)	Complete (B→R)	Δcomplete (pp)	ΔU
FaceNet	$+5.99$ $[+2.60, +9.38]$	$32 \to 47$	$+7.81$ $[+3.65, +11.98]$	$+0.024$ $[+0.015, +0.034]$
AdaFace	$+6.77$ $[+3.65, +9.90]$	$136 \to 159$	$+11.98$ $[+6.25, +17.71]$	$+0.054$ $[+0.046, +0.062]$

Strength sensitivity. Table 34 rescores existing Qwen outputs at four additional strengths on all 194 two-person and 33 three–four-person inputs. The two-person recall and completeness point estimates increase across these strengths, whereas three–four-person recall is non-monotonic. At $\alpha = 2.5$, the latter has five newly complete images and no regressions ($14 \to 19$); despite the positive percentile-bootstrap interval, the exact two-sided McNemar test gives $p = 0.0625$. These small-sample results do not establish a universally preferable strength, improved aesthetics, or multiplicity-adjusted significance.

Table 34: Qwen strength sensitivity versus baseline at $\mathrm{FPR} = 1\%$. p2: 194 two-person images; p3: 33 three–four-person images. Changes in recall and completeness are pp. Brackets are pointwise paired-bootstrap 95% CIs; $\alpha = 1.5$ is in Table 31.

Split	α	Δrecall (pp)	Complete (B→R)	Δcomplete (pp)	ΔU
p2	1.25	$+4.12$ $[+1.80, +6.44]$	$136 \to 149$	$+6.70$ $[+2.58, +10.82]$	$+0.036$ $[+0.029, +0.044]$
p2	1.75	$+5.93$ $[+3.09, +8.76]$	$136 \to 155$	$+9.79$ $[+5.15, +14.43]$	$+0.074$ $[+0.065, +0.083]$
p2	2.00	$+7.99$ $[+4.90, +11.08]$	$136 \to 161$	$+12.89$ $[+7.73, +18.56]$	$+0.090$ $[+0.080, +0.101]$
p2	2.50	$+11.60$ $[+8.25, +14.95]$	$136 \to 172$	$+18.56$ $[+12.89, +24.23]$	$+0.111$ $[+0.100, +0.122]$
p3	1.25	$+4.55$ $[+0.51, +9.09]$	$14 \to 16$	$+6.06$ $[-6.06, +18.18]$	$+0.038$ $[+0.016, +0.057]$
p3	1.75	$+3.79$ $[-1.52, +9.34]$	$14 \to 16$	$+6.06$ $[-9.09, +21.21]$	$+0.086$ $[+0.058, +0.112]$
p3	2.00	$+4.04$ $[-0.51, +9.09]$	$14 \to 17$	$+9.09$ $[-3.03, +24.24]$	$+0.098$ $[+0.067, +0.127]$
p3	2.50	$+7.83$ $[+2.78, +13.38]$	$14 \to 19$	$+15.15$ $[+3.03, +27.27]$	$+0.137$ $[+0.114, +0.162]$

The same existing-output audit at $\alpha = 2$ on FLUX retains all $194/33$ inputs and exactly reproduces the baseline calibration thresholds (Table 35). Two-person completeness has 41 gains and one regression; three–four-person completeness has eight gains and no regressions (exact two-sided McNemar $p = 0.0078125$ for the latter). These are within-backbone strength checks, not cross-backbone comparisons or evidence that the higher strength improves aesthetics.

Quality and completeness on identical samples. Table 36 pairs the archived, untrimmed AES scores with recall and U on the same IDs, using the intersection with valid AES in baseline and every tested strength within each backbone and split. These AES-eligible cohorts are distinct from the all-input cohorts above; the original calibration thresholds are held fixed. Stronger Qwen routing

Table 35: FLUX strength sensitivity at $\alpha = 2$ versus baseline, $\mathrm{FPR} = 1\%$, with all predefined eligible inputs retained. Recall and completeness changes are pp; brackets are pointwise 95% CIs.

Split	n	Δrecall (pp)	Complete (B→R)	Δcomplete (pp)	ΔU
p2	194	$+13.40$ $[+10.05, +16.75]$	$112 \to 152$	$+20.62$ $[+14.95, +26.80]$	$+0.093$ $[+0.084, +0.102]$
p3	33	$+13.64$ $[+7.32, +20.71]$	$13 \to 21$	$+24.24$ $[+12.12, +39.39]$	$+0.118$ $[+0.101, +0.136]$

shows negative AES intervals at $\alpha \geq 2$ for two-person and $\alpha \geq 1.75$ for three–four-person images. FLUX three–four-person AES decreases at both tested strengths, whereas its two-person intervals cross zero. Thus identity retention and aesthetic quality need not improve together. Intervals crossing zero do not establish aesthetic equivalence, non-inferiority, or Pareto dominance.

Table 36: Paired AES and completeness changes versus baseline on the same AES-eligible, across-strength common IDs. p2 denotes two-person and p3 three–four-person images. Recall changes are pp at $\mathrm{FPR} = 1\%$; AES and U retain their original scales. Brackets are pointwise, untrimmed paired-bootstrap 95% CIs ($10{,}000$ image resamples, PCG64 seed 20260818), without multiplicity correction.

Backbone	Split	α	n	ΔAES	Δrecall (pp)	ΔU
Qwen	p2	1.25	191	-0.004 $[-0.025, +0.018]$	$+3.93$ $[+1.57, +6.28]$	$+0.036$ $[+0.029, +0.044]$
Qwen	p2	1.5	191	-0.020 $[-0.045, +0.006]$	$+4.97$ $[+2.36, +7.85]$	$+0.057$ $[+0.050, +0.065]$
Qwen	p2	1.75	191	-0.022 $[-0.048, +0.005]$	$+5.76$ $[+2.88, +8.64]$	$+0.074$ $[+0.065, +0.083]$
Qwen	p2	2.0	191	-0.036 $[-0.066, -0.006]$	$+7.85$ $[+4.71, +11.26]$	$+0.090$ $[+0.079, +0.100]$
Qwen	p2	2.5	191	-0.043 $[-0.071, -0.015]$	$+11.52$ $[+8.12, +14.92]$	$+0.111$ $[+0.099, +0.122]$
Qwen	p3	1.25	31	-0.007 $[-0.057, +0.045]$	$+4.84$ $[+0.54, +9.68]$	$+0.040$ $[+0.016, +0.060]$
Qwen	p3	1.5	31	-0.014 $[-0.082, +0.057]$	$+5.11$ $[+0.54, +9.95]$	$+0.064$ $[+0.035, +0.090]$
Qwen	p3	1.75	31	-0.056 $[-0.106, -0.003]$	$+4.03$ $[-1.61, +9.68]$	$+0.086$ $[+0.057, +0.114]$
Qwen	p3	2.0	31	-0.085 $[-0.157, -0.008]$	$+4.30$ $[-0.54, +9.41]$	$+0.100$ $[+0.066, +0.129]$
Qwen	p3	2.5	31	-0.092 $[-0.166, -0.017]$	$+8.33$ $[+2.96, +14.25]$	$+0.140$ $[+0.115, +0.166]$
FLUX	p2	1.5	194	-0.015 $[-0.035, +0.005]$	$+7.99$ $[+4.90, +11.08]$	$+0.054$ $[+0.048, +0.061]$
FLUX	p2	2	194	-0.011 $[-0.036, +0.015]$	$+13.40$ $[+10.05, +16.75]$	$+0.093$ $[+0.084, +0.102]$
FLUX	p3	1.5	31	-0.055 $[-0.112, -0.002]$	$+5.11$ $[+0.81, +9.41]$	$+0.078$ $[+0.065, +0.091]$
FLUX	p3	2	31	-0.068 $[-0.126, -0.011]$	$+14.52$ $[+8.06, +21.77]$	$+0.120$ $[+0.101, +0.139]$

A stricter intersection requiring all six quality measures (AES, CLIP-I, CLIP-T, reference/ground-truth face similarity, and CP) retains only $164/20$ Qwen and $169/24$ FLUX p2/p3 images. Face similarities and CP average three backends only when all are available; missing values are not imputed. Their changes have positive pointwise intervals on these smaller cohorts, but AES conclusions are eligibility-sensitive: FLUX p2 at $\alpha = 1.5$ changes by -0.023 $[-0.045, -0.003]$ on 169 images, versus -0.015 $[-0.035, +0.005]$ on 194. These face similarities are not XVerseBench ID/IP scores. The archive retains metric-specific and joint cohorts, verifies per-sample raw/merged agreement and current staging/output byte identity, and checks source/data hashes before and after analysis. This establishes current archive consistency, not historical generation or evaluator-weight provenance.

Calibration diagnostics. Re-extracting the reference impostor distributions for Qwen two-person S0, three–four-person, and small-face reproduces every saved threshold exactly. Their primary thresholds are 0.3159, 0.1662, and 0.1902, based on 194, 141, and 31 within-sample impostor pairs, respectively. The archived diagnostics include raw similarities and 40-bin histograms on $[-1, 1]$ for impostors and assigned generated-reference pairs, retaining negative similarities and recording unassigned slots separately. Assigned pairs are selection-dependent, not ground-truth genuine pairs; these diagnostics do not rule out reference-to-generation domain shift or independently validate the target false-positive rate on generated faces.

Discordant-case inspection. A deterministic, regression-first selection yields 20 unique pairs from 117 eligible discordant cases (12 completeness regressions and 8 improvements). Non-blinded AI-assisted inspection finds clear foreground face-visibility or subject-count loss in two regressions; the other ten show no obvious such loss. None of the eight improvements visibly restores an absent person. One improved FLUX three–four-person case contains five visible people in both arms, illustrating that requested-identity coverage is not exact-count accuracy. Image hashes, selection rules, and descriptive observations are archived. This targeted inspection is neither a random sample

nor a human identity study: its counts are not population rates and do not identify the cause of embedding-score changes.

S MASS-MATCHED KEY-SCOPE CONTROL

Protocol. Equal $\log \alpha$ on different key sets does not inject equal attention mass, so the equal-α key-scope comparison in Appendix C confounds scope content with scope mass. We therefore regenerate three arms under one pinned two-person protocol: face-local routing at $\alpha = 1.5$, fixed whole-reference routing at $\alpha = 1.5$, and a mass-matched whole-reference arm D. Writing F_j and W_j for the unnormalized face and total mass of reference j at a given target query and head, D scales whole-reference keys by $1 + (\alpha - 1)F_j/W_j$. At these fixed logits, the normalized reference totals equal those obtained by applying the face-local rule to the same logits, while the selected scope remains whole-reference. Reference sets are disjoint, so the match holds per reference. Each arm can subsequently follow a different trajectory; this construction does not force their realized attention statistics to agree throughout generation. This protocol regenerates all three arms and is not pooled with the historical runs reported elsewhere; absolute values are therefore not comparable across protocols, only within this table.

Paired statistics follow Appendix C: exact common IDs, untrimmed per-sample means over ArcFace, FaceNet, and AdaFace, percentile bootstrap 95% intervals (10,000 replicates) and two-sided sign-flip tests (100,000 replicates, seed 20260913). Common-ID counts are 193 for Qwen and 194 for FLUX; CP denominators are smaller where a backend emits no CP.

Table 37: Mass-matched key-scope control on the two-person split under one pinned protocol. Upper block: arm means. Lower block: paired deltas with 95% intervals. Positive $\Delta\mathrm{M_{CP}}$ is a cost. Intervals crossing zero are unresolved and are not equivalence claims.

Backbone	Arm	$\mathrm{sim_{ref}}$	$\mathrm{sim_{GT}}$	$\mathrm{M_{CP}}$
QWEN-IMAGE-EDIT-2511	face-local ($\alpha = 1.5$)	0.5945	0.4147	0.2165
QWEN-IMAGE-EDIT-2511	whole-reference ($\alpha = 1.5$)	0.6033	0.4190	0.2232
QWEN-IMAGE-EDIT-2511	D (mass-matched)	0.5920	0.4133	0.2146
FLUX.2-KLEIN-BASE-9B	face-local ($\alpha = 1.5$)	0.4770	0.3719	0.1208
FLUX.2-KLEIN-BASE-9B	whole-reference ($\alpha = 1.5$)	0.4853	0.3794	0.1232
FLUX.2-KLEIN-BASE-9B	D (mass-matched)	0.4743	0.3709	0.1183

Backbone	Contrast	Metric	n	Δ [95% CI]	p
QWEN-IMAGE-EDIT-2511	whole − face	$\Delta\mathrm{sim_{ref}}$	193	$+0.0088$ $[+0.0013, +0.0166]$	0.0254
QWEN-IMAGE-EDIT-2511	whole − face	$\Delta\mathrm{sim_{GT}}$	193	$+0.0043$ $[-0.0008, +0.0094]$	0.1000
QWEN-IMAGE-EDIT-2511	whole − face	$\Delta\mathrm{M_{CP}}$	186	$+0.0082$ $[+0.0009, +0.0157]$	0.0322
QWEN-IMAGE-EDIT-2511	D − face	$\Delta\mathrm{sim_{ref}}$	193	-0.0025 $[-0.0068, +0.0018]$	0.2550
QWEN-IMAGE-EDIT-2511	D − face	$\Delta\mathrm{sim_{GT}}$	193	-0.0014 $[-0.0053, +0.0022]$	0.4618
QWEN-IMAGE-EDIT-2511	D − face	$\Delta\mathrm{M_{CP}}$	186	-0.0026 $[-0.0068, +0.0016]$	0.2248
QWEN-IMAGE-EDIT-2511	whole − D	$\Delta\mathrm{sim_{ref}}$	193	$+0.0113$ $[+0.0039, +0.0191]$	0.0028
QWEN-IMAGE-EDIT-2511	whole − D	$\Delta\mathrm{sim_{GT}}$	193	$+0.0057$ $[+0.0010, +0.0103]$	0.0166
QWEN-IMAGE-EDIT-2511	whole − D	$\Delta\mathrm{M_{CP}}$	185	$+0.0111$ $[+0.0035, +0.0192]$	0.0048
FLUX.2-KLEIN-BASE-9B	whole − face	$\Delta\mathrm{sim_{ref}}$	194	$+0.0083$ $[+0.0034, +0.0136]$	0.0016
FLUX.2-KLEIN-BASE-9B	whole − face	$\Delta\mathrm{sim_{GT}}$	194	$+0.0075$ $[+0.0033, +0.0118]$	0.0006
FLUX.2-KLEIN-BASE-9B	whole − face	$\Delta\mathrm{M_{CP}}$	178	$+0.0026$ $[-0.0021, +0.0072]$	0.2813
FLUX.2-KLEIN-BASE-9B	D − face	$\Delta\mathrm{sim_{ref}}$	194	-0.0027 $[-0.0058, +0.0003]$	0.0768
FLUX.2-KLEIN-BASE-9B	D − face	$\Delta\mathrm{sim_{GT}}$	194	-0.0010 $[-0.0035, +0.0016]$	0.4469
FLUX.2-KLEIN-BASE-9B	D − face	$\Delta\mathrm{M_{CP}}$	178	-0.0027 $[-0.0060, +0.0006]$	0.1132
FLUX.2-KLEIN-BASE-9B	whole − D	$\Delta\mathrm{sim_{ref}}$	194	$+0.0111$ $[+0.0065, +0.0159]$	0.00002
FLUX.2-KLEIN-BASE-9B	whole − D	$\Delta\mathrm{sim_{GT}}$	194	$+0.0085$ $[+0.0047, +0.0124]$	0.00002
FLUX.2-KLEIN-BASE-9B	whole − D	$\Delta\mathrm{M_{CP}}$	179	$+0.0049$ $[+0.0004, +0.0094]$	0.0388

Reading. Table 37 shows that the equal-α whole-reference identity advantage reproduces under this protocol on both backbones, and the Qwen CP cost of whole-reference routing reproduces as well. After mass matching, against D, face-local routing shows no resolved identity or CP difference on either backbone, while whole-reference routing at fixed α retains a resolved increase over D on all three metrics for both backbones, including the CP cost. These results identify injected reference mass as an important confound in the equal-α comparison; they do not isolate the identity contribution of non-face tokens or establish scope equivalence.

What this does not show. An unresolved interval is not evidence of equivalence; we pre-specified no equivalence margin, and these intervals are pointwise and uncorrected for the multiple contrasts in this table. D equalizes attention mass, not resulting identity similarity, so it does not answer what CP face-local routing would pay at matched $\mathrm{sim}_{\mathrm{ref}}$; that identity-matched comparison remains future work. An earlier variant that boosted only face keys to match whole-reference totals (arm C) failed its generation guards and is excluded from all inference here.

Preference-proxy reading of the same arms. The FLUX arms are also scored with a learned human-preference reward that never sees the reference images, and therefore lies outside the face-embedding space that defines $\mathrm{sim}_{\mathrm{ref}}$ and M_{CP} (Appendix T). The preference gap follows the same pattern as identity and CP: face-local routing scores above fixed whole-reference routing by $+0.2033$ $[+0.0777, +0.3389]$, but against D the contrast is unresolved ($+0.0129$ $[-0.0417, +0.0756]$) while whole-reference routing scores lower (-0.1903 $[-0.3114, -0.0773]$). This attenuation is consistent with injected mass contributing to the equal-strength reward gap; it does not exclude a role for scope content. This reward is a proxy rather than a human study, and this pinned protocol regenerates only the three routed arms, so it carries no unrouted baseline of its own; the separately scored cohorts that do include one rank every routed arm below it. See Appendix T for both caveats in full.

T Learned Preference Proxy

Protocol. Both M_{CP} and $\mathrm{sim}_{\mathrm{ref}}$ are built from the same face-embedding space, so they cannot by themselves establish that the cost of routing is visible outside that space. We therefore score the FLUX.2-KLEIN-BASE-9B arms with HPSv3 (Ma et al., 2025), a learned wide-spectrum human-preference reward. The scorer receives the complete generated image and its original generation prompt and never sees the reference images, so it is independent of the embeddings that define $\mathrm{sim}_{\mathrm{ref}}$ and M_{CP}. We report raw reward μ (higher is preferred) and paired contrasts over common sample IDs, with percentile bootstrap 95% intervals (5,000 replicates, seed 1729, pointwise and uncorrected). Images and prompts are deduplicated by content before scoring. Cohorts are scored and compared only within themselves; absolute values are not comparable across cohorts or protocols.

Results. At matched strength, face-local routing has higher mean HPSv3 reward than whole-reference routing in every tested cohort (Tables 38 and 39). The direct paired difference is resolved on P2 at both strengths ($+0.1839$ at $\alpha = 1.5$ and $+0.3171$ at $\alpha = 2.0$) and under the pinned protocol ($+0.2033$); the P3 and small-face intervals span zero. Under mass matching, face-local minus D is unresolved, while fixed whole-reference routing scores below D. This pattern is consistent with injected reference mass contributing to the equal-strength preference gap.

The unrouted baseline has higher mean reward than each routed arm in all three cohorts that include it. The baseline difference is resolved for whole-reference routing on P2 and P3 and for face-local routing at $\alpha = 2.0$, while the face-local $\alpha = 1.5$ intervals span zero on all three cohorts. Whole-reference routing's small-face interval also spans zero. On P2, the baseline-minus-arm gap increases from $+0.0773$ to $+0.1492$ for face-local routing and from $+0.2612$ to $+0.4663$ for whole-reference routing when α increases from 1.5 to 2.0.

The FLUX small-face reward cohort contains 33 images (32 for the crop contrast), distinct from the 49-example Qwen identity summary in Figure 3. Face-local routing differs from baseline by -0.0218 and exceeds cropped references by $+0.0393$; both paired intervals span zero.

Interpretation. HPSv3 provides a complementary image–prompt preference score, not a human vote or a direct measure of identity retention, background transfer, or aesthetics. These single-seed FLUX results use pointwise, uncorrected intervals; repeated identities across samples may reduce the

Table 38: HPSv3 arm means on FLUX.2-KLEIN-BASE-9B. Higher is preferred. Values are raw reward, not a rating on a fixed scale and not a human vote share.

Cohort	Arm	n	μ
two-person	unrouted baseline	194	14.6012
two-person	face-local ($\alpha = 1.5$)	194	14.5240
two-person	face-local ($\alpha = 2.0$)	194	14.4520
two-person	whole-reference ($\alpha = 1.5$)	194	14.3400
two-person	whole-reference ($\alpha = 2.0$)	194	14.1349
three-/four-person	unrouted baseline	33	14.9782
three-/four-person	face-local ($\alpha = 1.5$)	33	14.9597
three-/four-person	face-local ($\alpha = 2.0$)	33	14.6665
three-/four-person	whole-reference ($\alpha = 1.5$)	33	14.7439
three-/four-person	whole-reference ($\alpha = 2.0$)	33	14.4949
small-face	unrouted baseline	33	16.0019
small-face	face-local ($\alpha = 1.5$)	33	15.9801
small-face	whole-reference ($\alpha = 1.5$)	33	15.8609
small-face	cropped references	32	15.9093
pinned two-person (Appendix S)	face-local	194	14.5357
pinned two-person (Appendix S)	whole-reference	194	14.3325
pinned two-person (Appendix S)	D (mass-matched)	194	14.5228

Table 39: Paired HPSv3 contrasts within cohort, over common sample IDs, with percentile bootstrap 95% intervals (5,000 replicates, seed 1729). A positive baseline-minus-arm value is a preference cost of routing. Win counts are strict numeric comparisons of the reward without perceptual calibration. Intervals are pointwise and uncorrected, and an interval spanning zero is an unresolved contrast, not evidence of equivalence.

Cohort	Contrast	n	Δ [95% CI]	wins
two-person	baseline − face ($\alpha = 1.5$)	194	$+0.0773$ $[-0.0095, +0.1652]$	112/82
two-person	baseline − whole ($\alpha = 1.5$)	194	$+0.2612$ $[+0.1372, +0.3922]$	121/73
two-person	baseline − face ($\alpha = 2.0$)	194	$+0.1492$ $[+0.0391, +0.2572]$	120/74
two-person	baseline − whole ($\alpha = 2.0$)	194	$+0.4663$ $[+0.3116, +0.6309]$	132/62
two-person	face − whole ($\alpha = 1.5$)	194	$+0.1839$ $[+0.0632, +0.3132]$	116/78
two-person	face − whole ($\alpha = 2.0$)	194	$+0.3171$ $[+0.1734, +0.4716]$	121/73
three-/four-person	baseline − face ($\alpha = 1.5$)	33	$+0.0185$ $[-0.2400, +0.2388]$	21/12
three-/four-person	baseline − whole ($\alpha = 1.5$)	33	$+0.2343$ $[+0.0047, +0.4754]$	21/12
three-/four-person	baseline − face ($\alpha = 2.0$)	33	$+0.3117$ $[+0.0686, +0.5976]$	23/10
three-/four-person	baseline − whole ($\alpha = 2.0$)	33	$+0.4833$ $[+0.1820, +0.8178]$	24/9
three-/four-person	face − whole ($\alpha = 1.5$)	33	$+0.2158$ $[-0.0139, +0.5095]$	21/12
three-/four-person	face − whole ($\alpha = 2.0$)	33	$+0.1716$ $[-0.1307, +0.4999]$	17/16
small-face	baseline − face ($\alpha = 1.5$)	33	$+0.0218$ $[-0.0353, +0.0821]$	20/13
small-face	baseline − whole ($\alpha = 1.5$)	33	$+0.1410$ $[-0.0001, +0.3586]$	21/12
small-face	face − whole ($\alpha = 1.5$)	33	$+0.1192$ $[-0.0341, +0.3279]$	20/13
small-face	face − cropped references	32	$+0.0393$ $[-0.2344, +0.2885]$	21/11
pinned two-person	face − whole	194	$+0.2033$ $[+0.0777, +0.3389]$	113/81
pinned two-person	face − D	194	$+0.0129$ $[-0.0417, +0.0756]$	96/98
pinned two-person	whole − D	194	-0.1903 $[-0.3114, -0.0773]$	87/107

effective sample size. Scoring at roughly 448×448 may also limit sensitivity to fine facial detail. The measured reward differences therefore describe the tested image–prompt preference trade-off, with perceptual effects left to direct human evaluation.

U Appearance-Based Reference-Preference Audit

CP and identity similarity share recognition embeddings. We therefore repeat the reference-preference contrast with two hand-designed appearance descriptors. This changes the measurement of appearance, but not every stage of detection and matching.

Protocol. SCRFD detection and ArcFace-based reference assignment follow Appendix H. Assigned faces are warped to 112×112 grayscale crops by a five-point similarity transform, removing scale, in-plane rotation, and translation. We compute zero-mean normalized cross-correlation (intensity) and cosine similarity of a dense $7 \times 7 \times 9$ gradient-orientation histogram (texture). The descriptors have no trained weights; the upstream detector and identity assignment do. Reference preference is

$$\rho_d = \frac{d(g,t) - d(g,r)}{d(g,t) + d(g,r)}, \qquad d(a,b) = 1 - \mathrm{sim}(a,b). \tag{12}$$

We use 10,000 paired-bootstrap resamples with analysis seed 20260918.

Results. On 187 common Qwen two-person examples, texture preference increases from 0.0789 to 0.0925: $\Delta = +0.0137$ $[0.0053, 0.0220]$. Intensity preference changes from 0.0461 to 0.0566, with an unresolved contrast of $+0.0105$ $[-0.0001, 0.0219]$. Texture similarity to the reference increases by 0.0074 $[0.0031, 0.0118]$. Changes in ground-truth similarity are unresolved in both channels ($+0.0006$ and -0.0078). On the three-/four-person split ($n = 26$), texture preference increases by 0.0256 $[0.0040, 0.0488]$, from 0.1132 to 0.1388, and texture reference similarity rises by 0.0111 $[0.0036, 0.0186]$. All intensity contrasts remain unresolved.

Interpretation. The texture result supplies evidence of increased reference preference beyond the recognition descriptor used by CP. It does not make the audit fully embedding-independent, establish literal copying, or identify sharpening as the cause. Magnitudes from different descriptors are not directly comparable, and intervals spanning zero do not establish unchanged appearance. Together with the geometry diagnostic, these results qualify the identity gain with a measured appearance cost rather than certifying that copying is absent.

V Pre-Bias Face-Key Attention-Mass Diagnostic

Equation 5 describes a fixed-logit operation. To locate its operating regime, we instrument the 40-step Qwen editor on 24 two-person examples at $\alpha = 1.5$, without a mass cap. At each attention operation, the logger computes the pre-bias face-key mass S and its analytical uncapped transformation S'. Crucially, the inference trajectory already contains routing in earlier operations. These measurements are not an independent unrouted baseline trajectory paired with a routed trajectory.

For each of the 60 transformer layers, the saved diagnostic averages over heads, target-query rows, and denoising steps, yielding 1,440 sample–layer cells. The recorded S' averages the row-wise transformation; it is not the transformation of a layer-averaged S, nor a separately measured output mass. An implementation guard at numerically saturated rows can slightly alter the applied offset. The legacy step field in these summaries is not a resolved denoising-step index, so no step-specific distribution is inferred.

Table 40: Pre-bias mass and its analytical transformation along the routed Qwen trajectory. Values are percentages of attention mass; each observation is a sample–layer mean pooled over denoising steps.

Quantity	Mean	Q1	Median	Q3	Max
Pre-bias S	1.655	0.867	1.417	2.219	7.446
Analytical S'	2.303	1.222	1.998	3.056	9.970
$S' - S$	0.648	0.360	0.576	0.863	2.524

Table 40 summarizes the distribution. The median pre-bias mass is 1.42%, with interquartile range $[0.87\%, 2.22\%]$. At $\alpha = 1.5$, the analytical transformed mass has median 2.00%; the median increase

is 0.58 percentage points. The increase is at least 0.5 percentage points in 57.6% of cells and at least 25% relative to pre-bias mass in 99.4%. These descriptive summaries show that the offset changes mass at the observed operating points. Cells share inputs and layers, so they are not independent benchmark samples. A separate unrouted run with step-resolved logging is still needed to characterize the baseline distribution; the present diagnostic neither supplies that distribution nor identifies the layers or steps causing the final identity gain.

W TEXT-KEY PATHWAY CONTROL

Question and protocol. This diagnostic tests whether the identity gain follows from perturbing any conditioning pathway or specifically from increasing access to reference-face evidence. We retain the 40-step Qwen two-person protocol, all 194 inputs, seed S0, $\alpha = 1.5$, target-image queries, model weights, and input images. The control changes only the biased key set: instead of adding $\log \alpha$ to reference-face keys, it adds the same offset to all text-token keys. It is a mechanistic negative control, not a task-matched competing method.

We intersect exact common sample IDs, compute untrimmed paired changes, and use 10,000 percentile-bootstrap replicates with analysis seed 20260818. Identity values average ArcFace, FaceNet, and AdaFace per sample. The CP entry below is explicitly ArcFace-only; other quality metrics are paired per-image evaluator outputs.

Table 41: Pathway control on the MultiID-Bench two-person split. Each cell is the paired mean change from the same unrouted baseline with a 95% bootstrap interval. Identity and quality metrics use $n = 194$; ArcFace CP uses $n = 178$ for text-key routing and $n = 180$ for reference-face routing. Positive CP is a cost. Intervals crossing zero are unresolved.

Biased keys	Metric	n	Δ [95% CI]
Text tokens	$\mathrm{sim}_{\mathrm{ref}}$	194	-0.0475 $[-0.0574, -0.0373]$
	$\mathrm{sim}_{\mathrm{GT}}$	194	-0.0215 $[-0.0282, -0.0149]$
	ArcFace CP	178	-0.0374 $[-0.0468, -0.0283]$
	AES	194	$+0.0370$ $[+0.0027, +0.0722]$
	CLIP-I	194	-0.0023 $[-0.0058, +0.0012]$
	CLIP-T	194	$+0.0004$ $[-0.0012, +0.0019]$
Reference-face tokens	$\mathrm{sim}_{\mathrm{ref}}$	194	$+0.0451$ $[+0.0384, +0.0521]$
	$\mathrm{sim}_{\mathrm{GT}}$	194	$+0.0262$ $[+0.0211, +0.0313]$
	ArcFace CP	180	$+0.0319$ $[+0.0252, +0.0387]$
	AES	194	-0.0214 $[-0.0465, +0.0035]$
	CLIP-I	194	$+0.0043$ $[+0.0016, +0.0072]$
	CLIP-T	194	$+0.0008$ $[-0.0003, +0.0019]$

Reading. Text-key routing reverses both identity effects (Table 41): its $\mathrm{sim}_{\mathrm{ref}}$ and $\mathrm{sim}_{\mathrm{GT}}$ intervals lie below zero, while the corresponding reference-face intervals lie above zero. The text control also lowers ArcFace CP, consistent with weaker reference preference, and slightly raises AES; its CLIP changes are unresolved. Thus, under the tested protocol, a positive logit offset is not generically identity-improving. The result supports pathway specificity of the reference access intervention, rather than the stronger claim that insufficient reference access is the only cause of identity drift.

Scope. Equal $\log \alpha$ does not equalize injected attention mass across text and face key sets, whose cardinalities and learned roles differ. This control therefore isolates the destination of the same scalar intervention, not a mass-matched text-versus-face comparison. It also uses one backbone, the two-person split, and one fixed generation seed; we do not claim that text-key routing must reduce identity under every architecture or prompt distribution.

X EXPLORATORY SINGLE-PERSON VIDEO EXTENSION

This study asks whether reference-face reweighting can also help a frozen image-to-video model retain a single identity. It uses MiniMax-H3's FL2VA variant (MiniMax, 2026), not either image backbone. The video intervention is trajectory-guided and requires two passes, so it is not a direct transfer of the image method's all-target query rule.

X.1 TWO-PASS ROUTING PROTOCOL

For each conditioning image, an unmodified generation supplies a baseline video. SCRFD detections are assigned to conditioning identities using ArcFace; missing track positions are interpolated, the trajectory is smoothed, and coordinates are resampled onto the latent temporal grid. A second generation reuses the input, prompt, and seed, but adds $\log 1.5$ from target-video queries inside the baseline face trajectory to the corresponding conditioning-frame face keys. The full conditioning image remains available.

The token layout contains text, conditioning video, target audio, and target video streams. Routing acts only on the selected target-video queries and conditioning-face keys. The recorded key-box margin ratio is 0.15 and the query-track margin ratio is 0.35; the offset is constant across transformer layers and generated frames, starting at frame 1. Text, audio, and non-face video queries receive no direct bias. The baseline trajectory can diverge from the second generation, so spatial restriction does not guarantee that biased queries continue to follow the generated face. The model is guidance-distilled and has no separate negative CFG pass in this protocol.

We evaluate ten single-person inputs with seed 42. Each video has 124 frames at 24 fps (approximately 5.17 seconds, from a requested duration of 5 seconds). Portrait and landscape outputs use 512×768 and 768×512 canvases, respectively, with aspect-preserving containment and black padding. The prompt deliberately keeps the face visible; it does not test arbitrary head motion, occlusion, or scene changes:

> A continuous full-body, front-facing shot. The person walks straight forward toward the camera at a relaxed, natural pace, with a gentle natural smile and relaxed arm swings. The person keeps facing the camera for the entire video; the full face, both eyes, and facial features remain clearly visible at all times. The person never turns around, turns sideways, looks away, or leaves the frame. The camera smoothly tracks backward at the same speed, keeping the person's head and face centered and fully inside the frame while maintaining a stable full-body composition and consistent scale. Preserve the person's identity, hairstyle, clothing, and the original setting. Smooth realistic motion, stable framing, no cuts.

The recorded runtime uses Python 3.10, PyTorch 2.10 with CUDA 12.8, a Diffusers 0.41 development build, and H100 GPUs. The two-pass adaptation incurs an additional baseline generation and face-tracking stage; no controlled video runtime comparison is available.

X.2 EVALUATION AND RESULTS

We evaluate every generated frame with zero-based indices 1–123, excluding the conditioning frame 0. The 20 single-person videos therefore contribute 2,460 evaluated frames, but the independent comparison unit is the paired input video, not the frame. Evaluation uses SCRFD (`det_10g`, 640^2 detector input, threshold 0.5) and ArcFace (`buffalo_l/w600k_r50`). Per-frame Hungarian assignment uses no identity-similarity acceptance threshold and no evaluation-time temporal tracking. Identity is averaged over valid assigned faces and frames within each video, then equally across videos. Both arms have 100% valid-face coverage and face-count accuracy on the single-person cohort.

Video-averaged identity rises by 0.0268, with nine positive paired changes; last-frame identity rises by 0.0434, with eight positive changes (Table 42). Table 43 exposes the variation: one input loses average identity, and two lose last-frame identity despite positive video means. No inferential claim is based on treating the 2,460 frames as independent samples.

Table 42: Descriptive identity results on ten paired single-person H3 inputs. Last-frame identity evaluates frame 123 only, not a final temporal window. Counts of improvements refer to paired inputs, not individual frames.

Metric	Baseline	Routed	Mean change	Improved pairs
Video-mean ArcFace	0.3752	0.4019	+0.0268	9/10
Last-frame ArcFace	0.3257	0.3690	+0.0434	8/10

Table 43: Per-input H3 identity changes, routed minus baseline. Anonymous IDs index the ten single-person inputs; no sample is removed from the summary.

Input	Video-mean change	Last-frame change
01	+0.0391	+0.0721
02	+0.0316	−0.0876
03	+0.0414	+0.1155
04	+0.0214	+0.0718
05	+0.0383	+0.0362
06	+0.0231	+0.0550
07	+0.0011	−0.0176
08	−0.0231	+0.0788
09	+0.0796	+0.0888
10	+0.0153	+0.0206

Scope of the extension. These results describe single-person identity retention on ten H3 inputs with one seed and face-visible prompts under the two-pass protocol. Motion, temporal quality, prompt adherence, audio, and human preference remain separate evaluation targets.

Y Local MultiID-Bench System Comparisons

Main-text Table 1 reports fourteen configurations on each of the two-person and three-/four-person splits. All entries come from our local evaluation artifacts, not numbers transcribed from the cited papers. The same metric names and aggregation rule align reporting, but do not make the inference protocols or metric-valid sets identical. Each backbone's Base-to-Whole comparison controls the backbone, prompt, references, and generation seed. The appendix records the protocols, coverage, and provenance needed to interpret the main comparison.

Table 1 groups XVerse, MOSAIC, and WithAnyone as trained personalization methods: their personalization components are learned beyond backbone pretraining. PAF and the reference-bias variants operate without additional personalization training; we include the unmodified backbones in this group as reference points. All systems use pretrained backbones. Color highlighting marks the best and second-best aggregates within each split and training group, ranked before rounding.

Y.1 Generation Protocols

Table 44: Recorded local generation settings for Table 1. The $\alpha = 1.5$ Qwen and FLUX.2 Whole variants bias all reference VAE keys from all target queries, without reference isolation. The primary seed is shown here; additional backbone seed tests are reported separately. A shared seed number across different systems does not imply matched latent noise or sampling trajectories.

System	Backbone	Output	Steps / guidance	Seed
XVerse	FLUX.1-dev	1024^2	28 / 3.5	42
MOSAIC	FLUX.1-dev	512^2	28 / 3.5	42
WithAnyone	FLUX.1-dev	1024^2	25 / 4.0	1234
PAF variant	FLUX.1-dev	1024^2	28 / 3.5	42
Qwen Base / Whole	Qwen-Image-Edit-2511	1024^2	40 / 4.0	See text
FLUX.2 Base / Whole	FLUX.2-klein-base-9B	1024^2	40 / 4.0	42

XVerse uses its released adapter, full RGB references, and deterministic subject-word anchors. MOSAIC uses its released subject LoRA at scale 1.0, reference scale 0.9, and official reference layout. This MultiID-Bench MOSAIC run is 512^2; it is distinct from the two MOSAIC configurations on XVerseBench in Section 6.3. WithAnyone uses ID and SigLIP weights of 0.5. Seven of its two-person generation retries use a fallback face-detection scale. The local FaceNet and AdaFace reference helpers select the first detected reference face where upstream handling can retain multiple detections. These compatibility changes preclude a claim that all preprocessing exactly reproduces the official pipeline. The primary Qwen seed is 1091359629774735.

What the PAF comparison tests. We run the released FLUX.1-dev PAF variant (Feng et al., 2026), not a full reproduction of every configuration in that paper. The source revision is `30e3b773a7723e525551bba65b77e49f7328b61b`. The recorded settings are $\tau = 0.6$, $\gamma = 1$, $\eta = 1$, maximum sequence length 512, concept processing enabled, and automatic perturbation disabled. Reference-person masks are obtained with SAM3; baseline-person masks come from previously generated FLUX.2 Base images. ArcFace Hungarian matching associates references with baseline faces, and unique person instances must cover at least 40% of the corresponding face. Source masks are face-scale aligned and translated, then intersected with baseline person segmentation. No ground-truth image or ground-truth person position is used for generation.

This protocol gives PAF additional target-layout information from another model's output. It evaluates a released training-free system under that layout policy, not a same-input or same-backbone comparison with the FACEROUTE variants. Its measured identity scores are lower than the routed backbones' in this run, but the contribution of backbone, preprocessing, layout extraction, and algorithm cannot be separated by these aggregates. Output person counts can differ from the requests; evaluation success alone does not measure count correctness or attribute all failures to PAF itself.

Y.2 METRIC SOURCES, DENOMINATORS, AND MISSING VALUES

For each identity metric, encoder-level collected scores are pooled before applying the evaluator's $1.5 \times$ IQR filter. The table uses `average.sim_ori.mean`, `average.sim_ref.mean`, `average.copy_geo.mean`, and `average.non_matched_ori.mean`. The last is Bld: non-matched *ground-truth* identity similarity, not the wrong-reference score `non_matched_ref`. CLIP-I, CLIP-T, and Aesthetic are the corresponding filtered general means; FID is a separately computed set-level statistic. The anatomy field is a placeholder and is not reported. CP measures relative embedding-space reference preference, not a literal copy rate.

Table 45: Supplementary FID and CP availability. A/F/D denote ArcFace, FaceNet, and AdaFace. Counts are samples per encoder before IQR filtering, not faces or independent replicates. FID uses the available evaluation images; the small three-/four-person split makes its estimates unstable.

	Two-person		Three-/four-person	
System	CP $n_{\mathrm{A}}/n_{\mathrm{F}}/n_{\mathrm{D}}$	FID ↓	CP $n_{\mathrm{A}}/n_{\mathrm{F}}/n_{\mathrm{D}}$	FID ↓
XVerse	169/164/164	114.64	26/27/27	186.19
MOSAIC	171/166/166	121.73	21/22/22	191.16
WithAnyone	172/166/166	113.08	28/29/29	173.75
PAF†	68/70/70	119.50	5/5/5	188.39
Qwen Base	181/180/180	115.18	23/23/23	186.27
Qwen + Whole image	185/179/179	114.40	24/24/24	186.82
FLUX.2 Base	177/173/173	118.35	27/27/27	195.34
FLUX.2 + Whole image	178/172/172	118.95	27/27/27	192.73

Processed records and metric availability. Processed/Total counts saved evaluation records over the benchmark size, 194 or 33; it does not count correctly generated people. Base and Whole retain completed 33-record P3 evaluations. Two ground truths missed by earlier JPG-only staging were recovered from their original PNG and WebP files and evaluated against the existing outputs. XVerse, MOSAIC, and WithAnyone also retain 33 P3 records. PAF retains 31: two four-identity inputs failed layout preparation before generation because their baseline-derived layouts had only three detected faces. All P2 configurations retain 194 records. Availability and IQR retention remain metric-specific; null values are not imputed.

Table 46: FACEROUTE-W minus Base at $\alpha = 1.5$. Untrimmed exact-ID paired identity effects with 95% percentile-bootstrap intervals (10,000 resamples). P2/P3 identity comparisons use 194/33 examples.

Backbone / split	$\Delta\mathrm{sim}_{\mathrm{ref}} \uparrow$	$\Delta\mathrm{sim}_{\mathrm{GT}} \uparrow$
Qwen / P2	$+0.0565$ [+0.0471, +0.0658]	$+0.0305$ [+0.0246, +0.0366]
Qwen / P3	$+0.0737$ [+0.0520, +0.0969]	$+0.0237$ [+0.0109, +0.0371]
FLUX.2 / P2	$+0.0470$ [+0.0400, +0.0544]	$+0.0292$ [+0.0235, +0.0351]
FLUX.2 / P3	$+0.0622$ [+0.0479, +0.0763]	$+0.0339$ [+0.0240, +0.0444]

WithAnyone's original two-person evaluation omitted three complete records when no encoder returned a reference-identity score. We reevaluate only these three outputs and retain their null identity metrics and measurable CLIP and Aesthetic scores, reusing the original 191 records unchanged. CLIP and Aesthetic aggregates and set-level FID are recomputed on the completed set; 194/194 denotes processing coverage, not 194 valid face-identity measurements. No images are regenerated or nulls imputed.

The main table reports both routing scopes at the indicated strengths. The paired scope comparison at $\alpha = 1.5$ uses 194 P2 and 33 P3 inputs from the same original generation arms. Multi-seed cohorts remain tied to their original arms.

Whole–Base paired identity effects. We average available encoder scores within each example, retain exact IDs with defined scores in both arms, and bootstrap the untrimmed paired differences 10,000 times (analysis seed 20260923). Generation seeds remain those in Table 44. Each identity comparison has 194 P2 or 33 P3 examples. The completed P3 scores reuse the same generated images and add only the two previously unstaged benchmark records.

Interpreting sparse CP. PAF CP is available for 68, 70, and 70 two-person samples, and only five three-/four-person samples per encoder before filtering (Table 45). These conditional values cannot establish less copying across the requested dataset. A low Bld likewise does not by itself establish identity preservation: a face may have low similarity to both the correct and incorrect identities. WithAnyone's $\mathrm{sim}_{\mathrm{GT}}$, Whole's identity gains, and the backbone Aesthetic scores therefore describe different aspects of system behavior, not one total ranking.

Verification boundary. A deterministic export records each summary's path and SHA-256 digest in the local audit manifest. Identity, CP, CLIP, and Aesthetic means are recomputed from serialized sample scores. Bld is reaggregated where per-sample values are available, including the completed Base evaluations. WithAnyone's historical Bld scores are recovered from ordered raw collectors after verifying the corresponding identity-score sequences and exactly reproducing the archived aggregates; only its three omitted records are newly evaluated. Other older JSON files that omit Bld use hashed summaries. FLUX FACEROUTE-W at $\alpha = 1.5$ uses the completed 194-image P2 Bld evaluation: 582 encoder scores before IQR filtering and 545 retained, giving 0.0746. Table 1 marks PAF's sparse CP cells explicitly. Completion audits verify exact ID sets, explicit null fields, original image hashes, and regression against historical per-sample scores. FID is recomputed on each completed set and otherwise taken from the archived summary; it is not averaged from per-sample values. Metric formulas and IQR filtering are unchanged, but local null retention and the reference-face helpers above preclude claiming end-to-end upstream equivalence. The export preserves all available precision and rounds only for display. A common-ID cross-system evaluation, a zero-penalty all-input comparison, and a harmonized end-to-end runtime comparison are not claimed here.

Y.3 PUBLISHED ATTENTION-CONTROL INTERFACES

Table 47 separates interface compatibility from empirical performance. PAF in Table 1 adds an actual training-free system comparison; it does not substitute for matched adaptations of Prompt-to-Prompt, MasaCtrl, or FreeControl to our frozen joint-attention backbones.

Table 47: Interface comparison for the published methods. Checks denote properties, not performance or impossibility of adaptation. A spatial condition means an external structure input, not an automatically derived foreground mask.

Method	No test-time training	Native multi-reference input	No external spatial condition	Existing ref.-VAE-key bias
Prompt-to-Prompt (Hertz et al., 2023)	✓	×	✓	×
MasaCtrl (Cao et al., 2023)	✓	×	✓	×
FreeControl (Mo et al., 2024)	✓	×	×	×
FACEROUTE	✓	✓	✓	✓