MIND THE REFGAP: CORRECTING REFERENCE ATTENTION IN DIFFUSION-BASED VISUAL EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reference-guided diffusion editors struggle to faithfully reproduce user-provided references. We identify a potential bottleneck in diffusion editors: many methods provide limited reference-attention allocation. For example, in LoomVideo, edit-region queries assign less than 1% of their attention mass to the reference. We introduce RefGAP, a training-free correction that determines logit-offset magnitudes online at each layer from the reference-attention mass measured during the forward pass. Positive offsets to reference logits strengthen reference usage by edit-region queries, while negative offsets for keep-region queries limit reference-induced changes outside the edit. Two global coefficients control the correction; they are selected once on validation data from four development diffusion editors and held fixed. Across seven diffusion-based image/video editors, RefGAP improves identity fidelity in head swapping and face swapping. RefGAP achieves a fidelity-preservation trade-off comparable to separately tuned constant edit-side biases, without per-approach strength sweeps. Additional experiments on virtual try-on and background replacement evaluate transfer beyond identity editing.

1 INTRODUCTION

Modern diffusion editors can condition on a user-provided reference image to edit a driving image or video, supporting tasks such as head swapping, face swapping, virtual try-on, and background replacement. Many recent systems adopt an *in-context* design (Jiang et al., 2025; Wu et al., 2025a;b; Black Forest Labs, 2025), in which reference tokens share the denoiser's attention sequence with the content being edited. Yet direct access to the reference does not guarantee fidelity.

Existing methods strengthen reference conditioning through task-specific training (Wang et al., 2025; Chen et al., 2024; Fang et al., 2024) or training-free manipulation of reference attention or features (Shin et al., 2025; Fan et al., 2024). However, inference-time interventions typically rely on fixed, empirically chosen coefficients that may require retuning across models or tasks. They do not explicitly couple reference-attention enhancement in edit regions with suppression in keep regions.

We partition each attention layer's keys into reference and rest tokens and compute the fraction ρ of each query's attention mass assigned to the reference (fig. 2). Using an edit mask, we average ρ separately over content queries in the edit and keep regions. The profiles reveal two patterns: some approaches exhibit substantially higher reference attention in edit versus keep regions within a localized layer band, typically in the network's second half; others assign little reference attention to edit queries across many layers, falling below 1% on LoomVideo (Wu et al., 2026). These patterns motivate selective layer intervention, while ρ provides an online signal for correction strength.

Based on this measurement, we introduce RefGAP, a training-free correction named after the reference-attention gap it aims to bridge. Within a single forward pass, using statistics already produced by the attention computation, RefGAP reads the reference-attention masses in both regions and adds one scalar to every reference-key logit: a positive offset for edit-region queries, so they read the reference more, and a negative offset for keep-region queries, so the stronger reference signal does not pull the preserved content with it. The offset adapts to the measured reference-attention mass at each intervened layer and forward pass, reducing the need for model- or task-specific tuning.

We calibrate two global coefficients and establish a shared layer-selection rule on four development approaches (JoyAI-Video-Edit (Xiao et al., 2026), LoomVideo (Wu et al., 2026), FLUX.2-klein and

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.

FLUX.2-klein-base (Black Forest Labs, 2025)) using 40 head-swap validation clips. We freeze the coefficients, thresholds, and rule before applying them to three held-out approaches. Each approach requires one validation profiling run to determine its layer set under the fixed rule; this set is reused across tasks without tuning. Identity similarity improves on head-swap and face-swap tasks. Without task-specific retuning, virtual try-on fidelity improves across seven approaches, while background replacement yields mixed results, including a clear degradation on Qwen-Image-Edit (section 5).

To summarize, our contributions are:

- We measure reference-attention mass for edit and keep queries, revealing approach-dependent allocation patterns and limited reference attention in several approaches.
- We propose RefGAP, a training-free, two-sided attention correction that derives its per-layer strength online from the measured attention mass. RefGAP increases reference influence inside the edit while suppressing reference attention in the keep region.
- We demonstrate cross-approach and cross-task transfer using a single configuration calibrated once on 40 head-swap validation clips. Without retuning, it consistently boosts identity fidelity across seven approaches in head and face swapping (alongside an attribute-preservation trade-off), transfers to virtual try-on, and delineates transfer boundaries with mixed results on background replacement.

2 Related Work

Reference-conditioned generation and editing. Reference-guided editors incorporate visual conditions either as tokens in a shared attention sequence or through a separate conditioning branch. In-context systems such as VACE (Jiang et al., 2025), LoomVideo (Wu et al., 2026), JoyAI-Video-Edit (Xiao et al., 2026), Qwen-Image-Edit (Wu et al., 2025a), OmniGen2 (Wu et al., 2025b), and FLUX.2-klein (Black Forest Labs, 2025) allow reference and content tokens to interact within joint attention. Decoupled methods inject reference features through cross-attention or additive branches, as in IP-Adapter (Ye et al., 2023). RefGAP targets the in-context setting, where reference and content tokens compete within the same softmax and their relative mass can be modified directly.

Training-free reference and attention steering. Training-free diffusion methods manipulate attention to control conditioning and source consistency (Hertz et al., 2022; Cao et al., 2023). Diptych Prompting (Shin et al., 2025) enhances attention from queries in the generated panel to keys in the reference panel. FreeCustom (Ding et al., 2024) combines multi-reference self-attention with weighted concept masks to integrate and emphasize reference concepts for multi-concept composition. RefDrop (Fan et al., 2024) controls consistency across generated images or video frames by mixing reference and self-attention outputs with a user-specified scalar shared across queries.

These methods control reference influence through fixed or user-specified coefficients selected for a particular model or task. FreeCustom and RefDrop primarily target generation or consistency, without defining a region to preserve in an existing source image. Diptych also supports subject-driven editing through an edit mask, but its reference-attention boost remains uniform across queries in the edited panel. GRAG (Zhang et al., 2026) and DCAG (Li, 2026) provide separate representation-level controls for editing strength through externally selected coefficients.

RefGAP determines logit-offset magnitudes online from reference attention measured separately for edit and keep queries. Positive offsets enhance reference attention in edit regions, while negative offsets suppress it in keep regions. After a single calibration, the same coefficients transfer to held-out approaches and additional tasks without further manual tuning.

3 Method

3.1 Overview

RefGAP defines the masked area as the edit region and its complement as the keep region. At each attention layer and denoising step, it measures reference-attention mass in both regions (section 3.2) to determine reference-logit shifts (section 3.3), applied according to a layer-selection rule (section 3.4). Figure 1 summarizes this inference-time pipeline, which requires no weight updates.

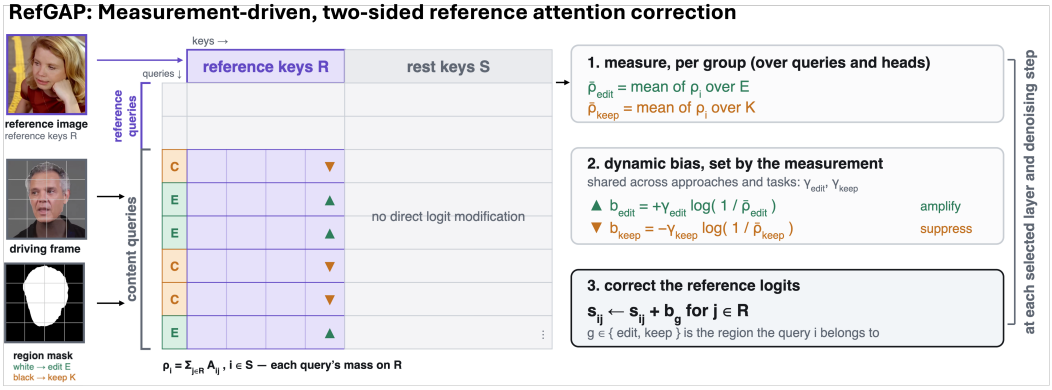


Figure 1: **RefGAP overview.** Attention-guided corrections strengthen reference attention inside the edit and suppress it outside, using two coefficients shared across approaches and tasks.

3.2 Reference attention mass and diagnosis

Consider attention layer ℓ at denoising step t, with H attention heads indexed by h. Let $s_{ij}^{(\ell,t,h)}$ denote the pre-softmax score between query i and key j in head h. We partition the keys into reference tokens $\mathcal{R}$ and all rest tokens $\mathcal{S}$. The *reference attention mass* of query i is the fraction of its attention allocated to the reference group:

$$\rho_i^{(\ell,t,h)} = \frac{\sum_{j\in\mathcal{R}} \exp(s_{ij}^{(\ell,t,h)})}{\sum_{j\in\mathcal{R}\cup\mathcal{S}} \exp(s_{ij}^{(\ell,t,h)})}, \qquad i \in S, \qquad \rho_i \in (0,1). \tag{1}$$

Let the edit mask, mapped to the latent token grid, partition content queries into an *edit set* $\mathcal{E}$ of tokens to be modified and a *keep set* $\mathcal{K}$ of tokens to be preserved. We average the per-head mass over the relevant queries and H heads, yielding region-level statistics per layer and denoising step:

$$\bar{\rho}_{\text{edit}}^{(\ell,t)} = \frac{1}{|\mathcal{E}|H} \sum_{i\in\mathcal{E}} \sum_{h=1}^{H} \rho_i^{(\ell,t,h)}, \qquad \bar{\rho}_{\text{keep}}^{(\ell,t)} = \frac{1}{|\mathcal{K}|H} \sum_{i\in\mathcal{K}} \sum_{h=1}^{H} \rho_i^{(\ell,t,h)}. \tag{2}$$

The first quantity measures how strongly queries in the edit region attend to the reference, while the second measures reference influence on queries in the keep region. Each statistic is a scalar for a given input, layer, and denoising step, recomputed from the current denoising trajectory immediately before that layer's offset is applied. This online computation requires no additional base-model forward pass; the offline profiling used for layer selection is described in section 3.4.

Token-count-normalized diagnosis. To account for reference-token count, we normalize reference-attention mass by the share assigned under uniform attention:

$$\pi_{\mathcal{R}} = \frac{|\mathcal{R}|}{|\mathcal{R}| + |\mathcal{S}|}, \qquad u_g^{(\ell,t)} = \frac{\bar{\rho}_g^{(\ell,t)}}{\pi_{\mathcal{R}}}, \quad g \in \{\text{edit}, \text{keep}\}. \tag{3}$$

Values below one indicate less reference attention than uniform allocation. This is a diagnostic baseline, not a target for the correction.

We average base-model statistics over all layers, denoising steps, and head-swap validation clips. Edit-region reference attention falls below the baseline in five approaches, with VACE the exception (table 1). JoyAI-Video-Edit is excluded because its reference-token share varies with the streaming KV cache. Qwen-Image-Edit and OmniGen2 assign more reference attention to keep than edit queries, motivating separate regional measurement and correction. Diagnostics appear in section B.

3.3 Two-sided reference correction

To control reference-attention allocation, RefGAP adds a shared offset to the reference-key logits for each query group. This changes the total attention assigned to the reference while preserving the

	LoomVideo	FLUX.2-klein	FLUX.2-klein-base	VACE	Qwen-Image-Edit	OmniGen2
u_{edit}	0.22	0.35	0.22	1.79	0.30	0.24
u_{keep}	0.13	0.10	0.15	0.69	0.35	0.28

Table 1: **Token-count-normalized reference attention.** Measured using head-swap validation data.

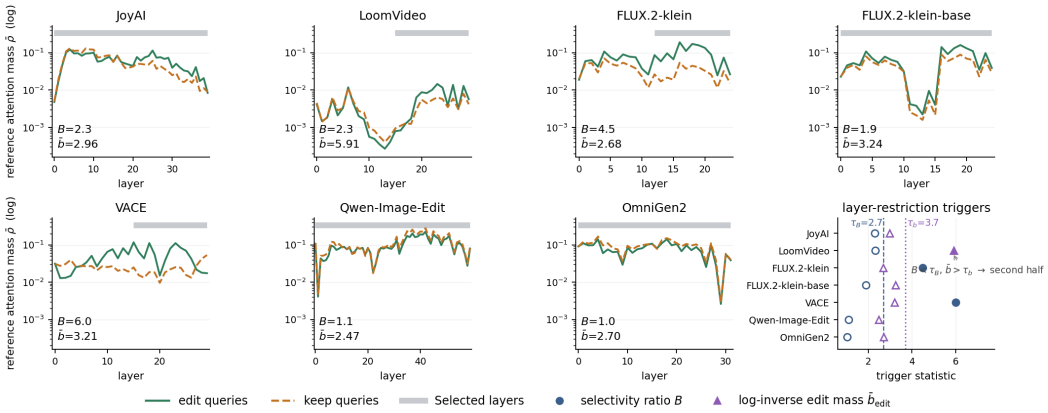


Figure 2: **Reference-attention profiles and layer selection.** Reference-attention masses across layers on the head-swap validation split, shown on a log scale. Gray bars mark layers selected by eq. (8). The bottom-right panels report complementary statistics B and $\bar{b}_{\text{edit}}$, with their thresholds.

relative attention weights within each key group. The offset is positive for edit queries and negative for keep queries; its magnitude adapts to the corresponding measured reference mass.

For query i in layer ℓ at step t, RefGAP adds a region-dependent offset to logits with keys in $\mathcal{R}$:

$$\tilde{s}_{ij} = s_{ij} + b^{(i)}_{\ell,t}, \qquad j \in \mathcal{R}, \tag{4}$$

with

$$b^{(i)}_{\ell,t} = \begin{cases} +\gamma_{\text{edit}} \log\left(1/\bar{\rho}^{(\ell,t)}_{\text{edit}}\right), & i \in \mathcal{E}, \\ -\gamma_{\text{keep}} \log\left(1/\bar{\rho}^{(\ell,t)}_{\text{keep}}\right), & i \in \mathcal{K}. \end{cases} \tag{5}$$

For edit queries, lower reference mass yields a larger positive offset, strengthening reference attention. For keep queries, we use a heuristic negative offset whose magnitude decreases as reference mass increases. This design aims to accommodate potentially necessary changes beyond the mask, such as hair extending past its boundary; reference mass alone does not distinguish these changes from unwanted leakage.

The coefficients γ_{edit} and γ_{keep} are calibrated once during development and shared across approaches and tasks. Applying the same offset to every reference-key logit for a query changes total reference mass while preserving relative weights within the reference and non-reference groups.

3.4 Layer Selection

We derive the layer-selection rule from two recurring patterns in the reference-attention profiles of the four development approaches. Figure 2 shows profiles and selected layers for seven approaches. First, some approaches exhibit clear edit-keep separation: $\bar{\rho}^{(\ell)}_{\text{edit}}$ dominates $\bar{\rho}^{(\ell)}_{\text{keep}}$ within a localized layer band, typically in the network's latter half, as in FLUX.2-klein (Black Forest Labs, 2025). Second, some assign little reference mass to edit queries across many layers, with $\bar{\rho}^{(\ell)}_{\text{edit}}$ remaining around 10^{-2}, as in LoomVideo (Wu et al., 2026). The first pattern suggests a selective reference-reading band; the second indicates a risk of accumulating large corrections when intervening throughout the network.

We capture these behaviors with two complementary statistics. For each group $g \in \{\text{edit}, \text{keep}\}$, let $\bar{\rho}_g^{(\ell)}$ denote the corresponding region mass from eq. (2), averaged over denoising steps and validation clips. We define the edit-to-keep selectivity ratio

$$B = \max_{\ell} \frac{\bar{\rho}_{\text{edit}}^{(\ell)}}{\bar{\rho}_{\text{keep}}^{(\ell)} + \epsilon_B}, \qquad \epsilon_B = 10^{-6}. \tag{6}$$

A large B indicates at least one layer assigns substantially more reference attention to the edit region than to the keep region. To summarize the potential strength of the edit-side correction, we compute

$$\bar{b}_{\text{edit}} = \frac{1}{L} \sum_{\ell=1}^{L} \log \left(\frac{1}{\bar{\rho}_{\text{edit}}^{(\ell)}} \right). \tag{7}$$

Whereas B measures layer-wise selectivity, $\bar{b}_{\text{edit}}$ averages the log-inverse edit mass across layers and serves as a proxy for reference-amplification strength. It is used only for layer selection; the deployed offsets remain those computed online at each attention layer and denoising step by eq. (5).

Given thresholds τ_B and τ_b, we define the set of correction layers as:

$$\mathcal{L}_{\text{RefGAP}} = \begin{cases} \{\lfloor L/2 \rfloor + 1, \ldots, L\}, & B \geq \tau_B \text{ or } \bar{b}_{\text{edit}} \geq \tau_b, \\ \{1, \ldots, L\}, & \text{otherwise.} \end{cases} \tag{8}$$

The thresholds and decision rule are fixed using only the four development approaches. For each approach, including the three held-out ones, we compute B and $\bar{b}_{\text{edit}}$ from its base-model head-swap validation profiles and apply the fixed rule to select correction layers. The resulting layer set is reused unchanged across tasks. This requires approach-specific profiling but no approach-specific strength or threshold tuning. Layer-selection ablations are reported in table 5.

4 Experiments

4.1 Experimental setup

Approaches. We evaluate seven diffusion editors: JoyAI-Video-Edit (Xiao et al., 2026), LoomVideo (Wu et al., 2026), FLUX.2-klein and FLUX.2-klein-base (Black Forest Labs, 2025), VACE (Jiang et al., 2025), Qwen-Image-Edit (Wu et al., 2025a), and OmniGen2 (Wu et al., 2025b).

Tasks and datasets. Our primary evaluation focuses on head and face swapping. Head swapping uses the 1,040 test clips of HeadSwapBench (Wang et al., 2025), and face swapping uses the 1,000 official target-source pairs from the FaceSwap subset of FaceForensics++ (Rossler et al., 2019).

We further evaluate transfer beyond identity editing through virtual try-on and background replacement. Try-on uses 90 ViViD clips (Fang et al., 2024), with reference garments permuted within category. Background replacement uses 150 ground-truth videos from the HeadSwapBench test split as inputs, paired with 30 SUN397 scene references (Xiao et al., 2010). Across all four tasks, video approaches process video clips, while image approaches edit the middle frame of each clip.

Metrics and comparison protocol.

Identity editing. For head and face swapping, we report AdaFace identity similarity (Kim et al., 2022), keypoint error (Lugaresi et al., 2019), and LPIPS (Zhang et al., 2018).

For head swapping, Repl. denotes the percentage of outputs whose generated face is closer in identity to the reference than to the input subject. Identity similarity is measured against the ground-truth video, and keypoint error is measured against the corresponding ground-truth frames.

For face swapping, Repl. is the percentage of outputs whose nearest identity in the 978-identity FF++ gallery is no longer the input subject, and Retr. is the percentage whose nearest identity is the reference. VACE's masked driving video hides the input identity, so we omit Repl. for VACE

in both identity tasks. Identity similarity is measured against the reference face, whereas keypoint error is measured against the input (driving) video.

For the main results, LPIPS uses task-specific preservation targets. For head swapping, it is computed over the whole frame against the ground-truth video. For face swapping, it is computed outside the face mask against the input video.

Transfer tasks. For virtual try-on and background replacement, we report region-specific DINOv2 similarity (Oquab et al., 2023) to the reference image over the edited garment and background regions, respectively. LPIPS is computed outside the garment mask against the input for virtual try-on, and inside the person mask against the input for background replacement.

Common evaluation protocol. Base and $RefGAP$ use identical inputs, random seeds, and input-derived masks. No generated outputs are used during mask construction. All metrics are macro-averaged over clips, and identity comparisons use only clips with valid scores from both methods. Implementation details are provided in section A.

4.2 Parameter calibration and transfer

We jointly explore the two global coefficients and the layer coverage using only a 40-clip head-swap validation split evaluated on four development approaches: JoyAI-Video-Edit, LoomVideo, FLUX.2-klein, and FLUX.2-klein-base. Specifically, we enumerate

$$\gamma_{\text{edit}} \in \{0.4, 0.6, 0.8\}, \qquad \gamma_{\text{keep}} \in \{0, 0.4, 0.8\}, \qquad \mathcal{L} \in \{\mathcal{L}_{\text{all}}, \mathcal{L}_{\text{half}}\},$$

where $\mathcal{L}_{\text{all}}$ denotes correction at all layers and $\mathcal{L}_{\text{half}}$ denotes correction at the second half of the network. This results in $3 \times 3 \times 2 = 18$ candidate settings, allowing the coefficient and layer-coverage choices to be explored jointly.

For each combination of correction coefficients and layer coverage, we measure identity performance together with the change in non-edit LPIPS (LPIPS_{ne}). We retain settings satisfying

$$\Delta\text{LPIPS}_{\text{ne}}(RefGAP) = \text{LPIPS}_{\text{ne}}(RefGAP) - \text{LPIPS}_{\text{ne}}(\text{Base}) \leq 0.005,$$

and select the best admissible coefficient pair according to validation identity performance.

In parallel, we validate whether the threshold-based rule correctly selects between all-layer and second-half correction. We set $\tau_B = 2.7$ and $\tau_b = 3.7$ to the arithmetic means of B and $\bar{b}_{\text{edit}}$, respectively, across the four development approaches. For each approach, we apply these fixed thresholds to its statistics and compare the resulting coverage with the corresponding all-layer and second-half results from the joint sweep. Full results are reported in table 8.

After this joint validation, we fix $\gamma_{\text{edit}} = 0.6$, $\gamma_{\text{keep}} = 0.8$, $\tau_B = 2.7$, and $\tau_b = 3.7$. We apply the shared rule to each approach's validation profile and reuse the selected layers across test sets and tasks, including the three held-out approaches. No further tuning is performed for these evaluations.

4.3 Results across approaches and tasks

Table 2 summarizes quantitative results across the seven approaches and four tasks, while fig. 3 presents qualitative examples of head swap, face swap, virtual try-on, and background replacement on three approaches (all seven in fig. 5).

Identity editing: head and face swapping. For head swapping, RefGAP raises the replacement rate to at least 86% on every applicable approach, improves identity similarity on all seven approaches, and does not increase whole-frame LPIPS on any approach, reducing it on four with intervals excluding zero (table 14). Keypoint error increases mainly on JoyAI-Video-Edit, FLUX.2-klein, and FLUX.2-klein-base. Section G.2 shows that delaying correction to later denoising steps substantially reduces keypoint degradation while retaining identity gains over the base model.

For face swapping, RefGAP improves identity similarity across all approaches, with replacement and correct retrieval rates reaching at least 82% and 75%, respectively. Keypoint error rises on most approaches, while non-face LPIPS varies by approach. Without ground-truth outputs, LPIPS compares against the input video, potentially counting identity-related geometry differences as preservation errors. Larger identity gains generally accompany larger keypoint changes (section G.1).

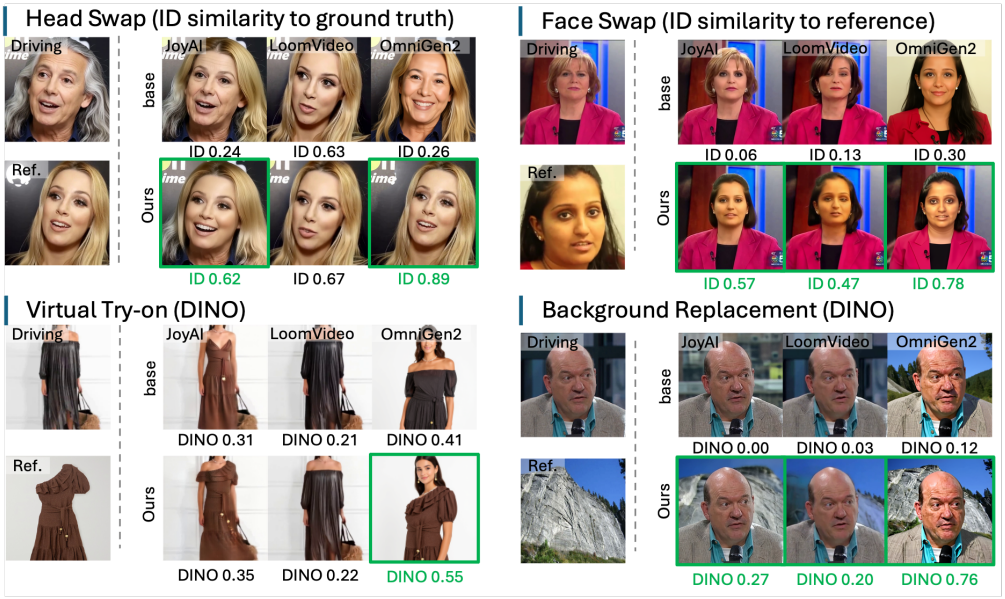


Figure 3: **Qualitative results across four tasks.** Base vs. Ours on JoyAI-Video-Edit (JoyAI), LoomVideo, and OmniGen2. Scores are shown below, with green borders marking gains ≥ 0.05.

	Head swap				Face swap					Virtual try-on		Background replacement	
Approach	Repl.↑	ID↑	Kpt.↓	LPIPS↓	Repl.↑	ID↑	Retr.↑	Kpt.↓	LPIPS↓	DINO↑	LPIPS↓	DINO↑	LPIPS↓
JoyAI	68→99% +31	.507→.644 +.137	.043→.085 +.042	.218→.215 −.003	60→97% +37	.282→.591 +.309	39→88% +49	.057→.105 +.048	.086→.108 +.022	.283→.298 +.015	.099→.101 +.002	.242→.286 +.044	.063→.065 +.002
LoomVideo	79→92% +13	.608→.665 +.057	.107→.098 −.009	.281→.275 −.006	71→84% +13	.449→.466 +.017	63→75% +12	.083→.086 +.003	.180→.178 −.002	.174→.183 +.009	.105→.105 .000	.135→.258 +.123	.051→.058 +.007
Klein	91→100% +9	.726→.790 +.064	.097→.153 +.056	.222→.217 −.005	98→100% +2	.704→.820 +.116	90→96% +6	.065→.082 +.017	.064→.076 +.012	.363→.372 +.009	.093→.092 −.001	.629→.630 +.001	.038→.037 −.001
Klein-base	77→100% +23	.588→.799 +.211	.169→.236 +.067	.326→.261 −.065	95→100% +5	.557→.831 +.274	74→96% +22	.129→.223 +.094	.173→.179 +.006	.409→.465 +.056	.150→.160 +.010	.632→.651 +.019	.162→.136 −.026
VACE	—	.405→.729 +.324	.198→.200 +.002	.228→.214 −.014	—	.541→.818 +.277	87→98% +11	.187→.193 +.006	.018→.019 +.001	.171→.226 +.055	.056→.056 .000	.008→.014 +.006	.008→.008 .000
Qwen	70→86% +16	.609→.723 +.114	.179→.179 .000	.449→.237 −.212	65→82% +17	.518→.687 +.169	59→79% +20	.097→.118 +.021	.069→.065 −.004	.213→.248 +.035	.077→.094 +.017	.299→.042 −.257	.027→.024 −.003
OmniGen2	34→99% +65	.207→.693 +.486	.255→.258 +.003	.464→.365 −.099	79→100% +21	.157→.656 +.499	20→91% +71	.238→.249 +.011	.323→.279 −.044	.277→.355 +.078	.422→.404 −.018	.245→.515 +.270	.259→.295 +.036
mean Δ	+26	+.199	+.023	−.058	+16	+.237	+27	+.029	−.001	+.037	+.001	+.029	+.002

Table 2: **Quantitative results across seven approaches and four tasks.** JoyAI, Klein, Klein-base, and Qwen denote JoyAI-Video-Edit, FLUX.2-klein, FLUX.2-klein-base, and Qwen-Image-Edit, respectively. Cells show Base→Ours, with differences shown below. Mean differences average all seven approaches (six for Repl., which is undefined for VACE). Green and red indicate improvements and degradations, respectively; gray denotes $|\Delta| < 0.005$. Differences are computed from the displayed values; unrounded paired estimates are given in table 14.

Transfer beyond identity editing. With the same coefficients, RefGAP improves virtual try-on reference fidelity on all seven approaches (table 2, intervals excluding zero on four). Background replacement has a different objective from identity editing and try-on: the surrounding scene should change, while the foreground subject should remain unchanged. It improves on JoyAI-Video-Edit, LoomVideo, and OmniGen2, changes little on the FLUX.2 variants and VACE, and degrades on Qwen-Image-Edit (table 14). These mixed results highlight that reference transfer remains task- and approach-dependent.

4.4 Human Evaluation

We evaluate 219 items with 54 raters retained after attention checks. Each item presents the input, reference, and base/RefGAP outputs in randomized A/B order. Raters assess reference fidelity, pose/expression consistency with the input, and visual quality, choosing A, B, both succeed, or

Approach	Head swap			Face swap			Virtual try-on			Background replacement		
	Ref.	Pose	Qual.	Ref.	Pose	Qual.	Ref.	Pose	Qual.	Ref.	Pose	Qual.
JoyAI-Video-Edit	28/0/0	3/25/0	5/23/0	50/7/0	3/50/4	5/50/2	37/5/12	0/52/2	5/46/3	19/25/6	0/50/0	1/46/3
LoomVideo	7/18/2	12/14/2	8/14/6	18/31/3	3/44/5	1/33/18	36/15/0	0/51/0	3/48/0	39/12/0	0/50/1	1/49/1
FLUX.2-klein	20/5/0	0/25/0	7/18/0	33/16/2	0/48/3	1/48/2	1/11/1	1/12/0	0/12/1	0/10/0	0/10/0	0/10/0
FLUX.2-klein-base	24/1/0	0/25/0	3/22/0	45/8/0	2/45/6	7/44/2	20/16/18	0/53/1	1/51/2	15/40/1	0/56/0	10/45/1
VACE	24/0/1	0/24/1	6/15/4	35/19/1	5/48/2	2/50/3	57/0/0	5/52/0	8/49/0	—	—	—
Qwen-Image-Edit	29/0/0	2/27/0	8/21/0	41/15/1	3/39/15	5/50/2	48/2/5	1/54/0	2/53/0	5/25/21	0/50/1	0/49/2
OmniGen2	25/0/0	10/15/0	0/25/0	56/0/1	3/49/5	11/45/1	50/1/4	2/53/0	5/50/0	49/9/1	5/53/1	5/51/2
All items	157/24/3	27/155/3	37/138/10	278/96/8	19/323/40	32/320/30	249/50/40	9/327/3	24/309/6	127/121/29	5/269/3	17/250/9

Table 3: **Human evaluation across four tasks.** Cells report judgment counts as RefGAP/tie/base for reference fidelity (Ref.), pose/expression consistency with the input (Pose), and visual quality (Qual.). Ties include "both succeed" and "both fail"; "—" denotes an unevaluated setting.

both fail. To prioritize informative comparisons, we use outputs with visible differences; results describe these selected comparisons. An unscreened audit of 56 randomly sampled comparisons by five raters favors RefGAP in reference fidelity (section F). Table 3 reports preferences for RefGAP in reference fidelity across head swap, face swap and virtual try-on, consistent with the generally improved identity similarity in head and face swapping and DINO scores in virtual try-on (table 2).

Preservation results are mixed. Pose/expression judgments are mostly ties (both-success or both-failure). Non-tied counts (RefGAP versus base) are 27 to 3 for head swapping and 19 to 40 for face swapping. Reference-fidelity judgments for background replacement vary across approaches. Quantitative results vary across approaches, including clear degradation on Qwen-Image-Edit.

4.5 Ablation studies

We ablate three RefGAP components: attention-conditioned strength, keep-side correction, and layer coverage, using the four development approaches on head swapping. We use non-edit LPIPS (LPIPS_{ne}) for validation and whole-frame LPIPS for the fixed-bias test comparison. Unlike the whole-frame LPIPS in table 2, LPIPS_{ne} is computed outside the dilated head mask to detect reference leakage into the keep region. For image approaches, validation metrics use the middle frame.

Approach	RefGAP			best admissible edit-side constant			
	ID↑	Kpt.↓	LPIPS↓	b	ID↑	Kpt.↓	LPIPS↓
JoyAI-Video-Edit	.644	.085	.215	1.5	.644	.075	.207
LoomVideo	.665	.098	.275	2.5	.671	.098	.271
FLUX.2-klein	.790	.153	.217	2.0	.794	.171	.229
FLUX.2-klein-base	.799	.236	.261	2.0	.801	.236	.260

Table 4: **RefGAP versus approach-specific constant edit-side biases.** Constants are selected on validation data under $\Delta\text{LPIPS}_{\text{ne}} \leq 0.005$. Only the edit-side bias differs between settings. Results use the full 1,040-clip head-swap test set; LPIPS is measured over the whole frame.

Online versus fixed strength. Diptych Prompting (Shin et al., 2025) boosts reference attention in the edited region by a uniform constant. As its editing code is unreleased, we compare against a constant-bias baseline inspired by Diptych Prompting: a constant edit-side bias b combined with our keep-side correction and layer sets, tuned separately for each approach. On the 40-clip validation split, we sweep $b \in \{0.5, 1, \ldots, 3\}$ and select the highest-identity value with $\Delta\text{LPIPS}_{\text{ne}} \leq 0.005$ (table 9). The selected constants range from 1.5 to 2.5, confirming that the optimal fixed strength is approach-dependent. On the 1,040-clip test set, they improve identity over RefGAP by only 0.000–0.006, with mixed keypoint and LPIPS trade-offs (table 4). RefGAP thus matches a per-approach-tuned Diptych-style baseline without per-approach strength sweeps.

One-sided versus two-sided correction. With $\gamma_{\text{edit}} = 0.6$, we vary $\gamma_{\text{keep}} \in \{0, 0.4, 0.8\}$, where zero gives the one-sided variant. Omitting keep-side correction severely degrades non-edit preservation (table 5, left): FLUX.2-klein-base's LPIPS_{ne} rises from 0.1206 (base) to 0.2115. Increasing γ_{keep} to 0.8 lowers LPIPS_{ne} across most models, reaching 0.0674 on FLUX.2-klein-base, while

Keep side: $\gamma_{\text{edit}} = 0.6$, rule-selected layers, $\gamma_{\text{keep}} \in \{0, 0.4, 0.8\}$							
		LPIPS$_{\text{ne}}\downarrow$			ID$\uparrow$		
Approach	base	$\gamma_{\text{keep}} = 0$	0.4	0.8	0	0.4	0.8
JoyAI-Video-Edit	.0663	**.0635**	.0644	.0641	.705	.705	.706
LoomVideo	.0700	.0781$^{\times}$	.0746	**.0727**	.768	.767	.767
FLUX.2-klein	.0435	.0524$^{\times}$	.0416	**.0406**	.885	.885	.885
FLUX.2-klein-base	.1206	.2115$^{\times}$	.0764	**.0674**	.879	.877	.878

layer coverage: $\gamma_{\text{edit}} = 0.6$, $\gamma_{\text{keep}} = 0.8$							
	layer rule			all layers		second half	
Approach	B	$\bar{b}_{\text{edit}}$	pred.	ID	LPIPS$_{\text{ne}}$	ID	LPIPS$_{\text{ne}}$
JoyAI-Video-Edit	2.30	2.96	all	**.706**	**.064**	.674	.067
LoomVideo	2.32	5.91	half	.826	.278	**.767**	**.073**
FLUX.2-klein	4.48	2.68	half	.893	.055	**.885**	**.041**
FLUX.2-klein-base	1.88	3.24	all	**.878**	**.067**	.826	.093

Table 5: **Keep-side and layer-coverage ablations.** Only γ_{keep} (left) or corrected layers (right) vary on the validation split. B and $\bar{b}_{\text{edit}}$ are computed from the base model's validation passes; pred. denotes the layer set selected by eq. (8). $^{\times}$ marks LPIPS$_{\text{ne}}$ increases > 0.005 over the base. Bold marks the lowest LPIPS$_{\text{ne}}$ (left) and the rule-selected setting (right).

largely preserving identity similarity. These results support keep-side correction as a means to contain strong edit-side interventions and limit reference-induced changes in preserved content.

Layer coverage. As specified in section 3.4, RefGAP applies the correction only to the second half of the network when $B \geq 2.7$ or $\bar{b}_{\text{edit}} \geq 3.7$, and to all layers otherwise. With $\gamma_{\text{edit}} = 0.6$ and $\gamma_{\text{keep}} = 0.8$, table 5 supports this rule. JoyAI-Video-Edit and FLUX.2-klein-base do not satisfy either threshold and perform better with all-layer correction. FLUX.2-klein is selected by its large B, and the second-half setting substantially lowers LPIPS$_{\text{ne}}$ from 0.055 to 0.041 with only a small ID decrease from 0.893 to 0.885. LoomVideo is selected by its large $\bar{b}_{\text{edit}}$: applying the correction to all layers gives higher ID (0.826) but increases LPIPS$_{\text{ne}}$ from the base value 0.070 to 0.278, whereas the second-half setting gives 0.767 ID and 0.073 LPIPS$_{\text{ne}}$. These results support using B to target selective reference-reading bands and $\bar{b}_{\text{edit}}$ to prevent excessive accumulation of large corrections.

5 Limitations

Identity–attribute trade-off. The all-step default prioritizes reference identity and can compromise pose and expression preservation. Delaying correction can mitigate these costs: on JoyAI-Video-Edit and FLUX.2-klein, skipping the first correction step substantially reduces keypoint error while retaining identity improvements over the base. However, the required delay varies across sampling schedules, and stronger geometric preservation reduces identity gains (table 19).

An approach-task-specific failure. Despite broad gains in identity editing and virtual try-on, RefGAP reduces background fidelity on Qwen-Image-Edit. Restricting layer coverage does not resolve this degradation, suggesting that this approach-task pair may require a different correction strategy.

Validation profiling overhead. Applying RefGAP to a new approach requires one base-model profiling run on a validation split to select correction layers under the fixed rule. The coefficients and thresholds remain unchanged, and the selected layers are reused across tasks. This adds setup cost, although it avoids repeated evaluations over a grid of correction strengths.

6 Conclusion

We introduce RefGAP, a training-free method that uses measured reference-attention mass to dynamically strengthen reference influence in edit regions and suppress it in keep regions. With two coefficients calibrated on four development approaches and a shared layer-selection rule, RefGAP improves identity fidelity in head and face swapping across seven approaches, including three held-out ones. It achieves a fidelity-preservation trade-off comparable to approach-specific constant edit-side biases without per-approach strength sweeps. Transfer without task-specific retuning yields positive fidelity gains in virtual try-on but mixed results in background replacement. Together, these results support measured attention as a practical signal for controlling reference influence, while leaving attribute preservation and reliable cross-task transfer as directions for further work.

AI USE STATEMENT

Generative AI tools were used to polish the writing, prepare comparison figures and user study materials, and check numerical consistency between tables and the underlying result files. The authors developed the method, designed the experiments, and interpreted the results.

ETHICS STATEMENT

This work studies reference-guided visual editing, including head and face swapping, which can be misused to fabricate media of real individuals. All images and videos used in our evaluation are drawn from public benchmarks (HeadSwapBench, FaceForensics++, ViViD, VITON-HD, and SUN397). We release no new identity data and neither train nor release a new model; our method is an inference-time modification of publicly released editors.

For the human evaluation, participants were undergraduate, master's, and doctoral students with a background in computer vision. All participants provided informed consent before taking part. We collected their judgments of the editing results without collecting personally identifiable information.

REPRODUCIBILITY STATEMENT

All seven approaches are publicly available and used through their released pipelines. RefGAP modifies attention logits at inference time and requires no training. The correction rule is described in section 3, and its coefficients and calibration procedure are reported in section 4.2. Section A provides implementation details and numerical safeguards; section C reports calibration results and additional ablations; section D describes task-specific evaluation protocols and additional results; and section F details the human evaluation and its statistical analysis. Code and per-clip result files will be released.

REFERENCES

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22503–22513. IEEE, 2023.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, 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, 2025. URL https://arxiv.org/abs/2511.16719.

Xuanhong Chen, Bingbing Ni, Yutian Liu, Naiyuan Liu, Zhilin Zeng, and Hang Wang. SimSwap++: Towards faster and high-quality identity swapping. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46 (1):576–592, 2024.

Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. VITON-HD: High-resolution virtual try-on via misalignment-aware normalization. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2021.

Ganggui Ding, Canyu Zhao, Wen Wang, Zhen Yang, Zide Liu, Hao Chen, and Chunhua Shen. FreeCustom: Tuning-free customized image generation for multi-concept composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jiaojiao Fan, Haotian Xue, Qinsheng Zhang, and Yongxin Chen. RefDrop: Controllable consistency in image or video generation via reference feature guidance. *arXiv preprint arXiv:2405.17661*, 2024.

Zixun Fang, Wei Zhai, Aimin Su, Hongliang Song, Kai Zhu, Mao Wang, Yu Chen, Zhiheng Liu, Yang Cao, and Zheng-Jun Zha. ViViD: Video virtual try-on using diffusion models, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

Minchul Kim, Anil K Jain, and Xiaoming Liu. AdaFace: Quality adaptive margin for face recognition. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 18729–18738. IEEE, 2022.

Guandong Li. Dual-channel attention guidance for training-free image editing control in diffusion transformers. *arXiv preprint arXiv:2602.18022*, 2026.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Mark McGuire, Jongmin Lee, Chuo-Ling Chang, Ming Guang Yong, Matthias Grundmann, and Vivek Kwatra. MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Art B. Owen. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2), December 2007. ISSN 1932-6157. doi: 10.1214/07-aoas122. URL http://dx.doi.org/10.1214/07-AOAS122.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–11, 2019.

Chaehun Shin, Jooyoung Choi, Heeseung Kim, and Sungroh Yoon. Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7986–7996. IEEE, 2025.

Yanan Wang, Shengcai Liao, Panwen Hu, Xin Li, Fan Yang, and Xiaodan Liang. DirectSwap: Mask-free cross-identity training and benchmarking for expression-consistent video head swapping. *arXiv preprint arXiv:2512.09417*, 2025.

Yingfeng Wang, Yuxuan Xiao, and Shengcai Liao. Head similarity: Modeling structured whole-head appearance beyond face recognition. *arXiv preprint arXiv:2605.07766*, 2026.

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, 2025a. URL https://arxiv.org/abs/2508.02324.

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

Jianzong Wu, Hao Lian, Jiongfan Yang, Dachao Hao, Ye Tian, Yunhai Tong, Jingyuan Zhu, Biaolong Chen, Qiaosong Qi, Aixi Zhang, Wanggui He, Mushui Liu, Pipei Huang, and Hao Jiang. LoomVideo: Unifying multimodal inputs into video generation and editing. *arXiv preprint arXiv:2606.06042*, 2026.

J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, June 2010. doi: 10.1109/CVPR.2010.5539970.

Yicheng Xiao, Wenxun Dai, Xinran Qin, Lin Song, Maoquan Zhang, Hang Xu, Yukang Chen, Yitong Li, Guohui Zhang, Yuan Zhang, Xuying Zhang, Tommy Zhang, Jianlong Yuan, Peihao Li, Shuai Lu, Siming Fu, Chuyang Zhao, Xin Han, Jie Huang, Wenbo Li, Guoqing Ma, Wei Huang, Xiaojuan Qi, Haoyang Huang, and Nan Duan. JoyAI-Video-Edit: Real-time open-ended video editing with autoregressive diffusion. *arXiv preprint arXiv:2608.03974*, 2026.

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.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Xuanpu Zhang, Xuesong Niu, Ruidong Chen, Dan Song, Jianhao Zeng, Penghui Du, Haoxiang Cao, Kai Wu, and An-an Liu. Group relative attention guidance for image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3840–3850, 2026.

APPENDIX

A IMPLEMENTATION AND STATISTICAL DETAILS

Default implementation settings. Unless otherwise specified in ablations or diagnostic experiments, we use eq. (5) with $(\gamma_{\text{edit}}, \gamma_{\text{keep}}) = (0.6, 0.8)$. Using zero-based indices, correction is applied to layers 15–29 on LoomVideo and VACE, and 12–24 on FLUX.2-klein (25 layers). All layers are corrected on JoyAI-Video-Edit (40), FLUX.2-klein-base (25), Qwen-Image-Edit (60) and OmniGen2 (32). The edit mask partitions content queries only; no additional mask is applied within the reference-key group. Base and RefGAP use identical pipeline settings and seeds, with the correction disabled for the base.

		Denoising (s/sample)			End-to-end (s/sample)			Peak memory	
Approach	Steps	Base	RefGAP	Δ	Base	RefGAP	Δ	Base (GB)	Δ (MiB)
VACE	30	158.2	159.0	+0.5%	159.4	160.0	+0.4%	26.30	+0
LoomVideo	50	106.2	108.9	+2.5%	112.0	116.6	+4.1%	43.95	+1.5
Qwen-Image-Edit	20	62.48	66.00	+5.6%	64.0	67.4	+5.3%	57.95	+0
OmniGen2	50	21.68	23.39	+7.9%	22.4	24.2	+8.0%	15.36	+0
FLUX.2-klein-base	50	16.35	20.34	+24%	17.0	21.2	+25%	15.55	+0
FLUX.2-klein	4	0.53	0.89	+68%	1.4	1.6	+14%	15.54	+75
JoyAI-Video-Edit[†]	2	—	—	—	12.2	18.8	+54%	55.9[‡]	+1400[‡]

Table 6: **Inference cost of RefGAP on head swapping.** Means over five timed samples after one warm-up, Base and RefGAP back to back on the same A100-80GB GPU with the deployed settings. Denoising time excludes encoding, decoding, and saving; end-to-end time excludes model loading. Peak memory is the PyTorch allocator's maximum allocation. [†]Unfused reference implementation that materializes the attention matrix in corrected layers; its streaming server reports no per-step timing. [‡]Whole-GPU peak from `nvidia-smi`.

Compute. Our fused implementation computes attention separately over reference and non-reference keys, then combines the outputs using log-sum-exp statistics that also yield the reference mass, without materializing the attention matrix. Table 6 reports costs for all seven approaches under the head-swap settings in table 2, including deployed layers, coefficients, resolutions, and step counts. Base and RefGAP run consecutively on the same otherwise idle A100-80GB GPU, with one warm-up clip followed by five timed validation clips. Denoising time comes from sampler logs; end-to-end time is measured between consecutive saved outputs, excluding model loading, at one-second resolution. Except for JoyAI-Video-Edit, peak memory is measured in-process as PyTorch's peak allocated bytes.

Fused denoising overhead is 0.5–7.9% on VACE, LoomVideo, Qwen-Image-Edit, and OmniGen2, and 24% and 68% on FLUX.2-klein-base and FLUX.2-klein, respectively. For the FLUX.2 variants, short image sequences make attention inexpensive relative to the additional kernel call and output combination; FLUX.2-klein's overhead is 0.37 s per image over four steps. Peak allocated memory is unchanged on four of these six approaches and increases by 1.5 MiB on LoomVideo and 75 MiB on FLUX.2-klein.

JoyAI-Video-Edit instead uses an unfused hook that materializes per-head fp32 attention matrices. This implementation incurs 54% end-to-end overhead and 1.4 GB additional peak memory. Because its serving process was not instrumented, memory is measured at the whole-GPU level using `nvidia-smi`, rather than PyTorch allocation statistics.

Numerical safeguards. For numerical stability, we floor both region-averaged reference-attention masses, $\bar{\rho}_{\text{edit}}$ and $\bar{\rho}_{\text{keep}}$, at 10^{-6} before computing the logarithms in the bias rule. We then clip the edit-side bias to $[0, 8]$; the keep-side bias is not clipped. The ϵ_B in eq. (6) stabilizes only the layer-selection ratio and is not used in computing the attention biases.

B REFERENCE-MASS DIAGNOSIS ACROSS APPROACHES

Table 7 supplements the normalized diagnostics in table 1 with absolute reference-attention masses and the layer-selection statistics B and $\bar{b}_{\text{edit}}$ for all seven base models on the head-swap validation split. The uniform reference-token share $\pi_{\mathcal{R}}$ and normalized masses u_g follow eq. (3). We omit these quantities for JoyAI-Video-Edit because its reference-token share varies with the streaming KV cache.

The mean log-inverse edit mass $\bar{b}_{\text{edit}}$ ranges from 2.47 to 5.91, indicating substantial variation in reference-attention allocation across approaches. This statistic summarizes the potential correction strength for layer selection; it is not the mean deployed offset, which is computed online using eq. (5).

C CALIBRATION AND ADDITIONAL ABLATIONS

C.1 CALIBRATION GRIDS

We jointly evaluate layer coverage $\{\text{all}, \text{second half}\}$, $\gamma_{\text{edit}} \in \{0.4, 0.6, 0.8\}$, and $\gamma_{\text{keep}} \in \{0, 0.4, 0.8\}$ on the head-swap validation split, yielding 18 configurations per development approach (table 8). The coefficients are selected jointly and shared across approaches and tasks.

Under the rule-selected layer sets, four coefficient pairs satisfy the preservation criterion and improve identity on all four approaches: $\gamma_{\text{edit}} \in \{0.4, 0.6\}$ and $\gamma_{\text{keep}} \in \{0.4, 0.8\}$. Increasing γ_{edit} to 0.8 degrades JoyAI-Video-Edit below its base identity score and violates preservation on JoyAI-Video-Edit and LoomVideo, while removing keep-side correction ($\gamma_{\text{keep}} = 0$) violates preservation at both remaining edit coefficients on FLUX.2 variants, and at $\gamma_{\text{edit}} = 0.6$ on LoomVideo.

Among the four feasible pairs, $\gamma_{\text{edit}} = 0.6$ gives higher identity similarity than $\gamma_{\text{edit}} = 0.4$ on every approach at either keep coefficient. At $\gamma_{\text{edit}} = 0.6$, increasing γ_{keep} from 0.4 to 0.8 lowers non-edit LPIPS on all four approaches. Identity changes by at most 0.001 on all four. We therefore select $(\gamma_{\text{edit}}, \gamma_{\text{keep}}) = (0.6, 0.8)$ as a shared fidelity-preservation compromise. The lower block reports transfer to approaches excluded from coefficient selection. These ablations support keep-side suppression but do not establish that its mass-dependent form is preferable to constant suppression.

C.2 ONLINE VERSUS FIXED EDIT-SIDE STRENGTH

We compare online correction with approach-specific fixed edit-side biases on the 40-clip head-swap validation split using the four development approaches. We sweep $b \in \{0.5, 1, 1.5, 2, 2.5, 3\}$ while keeping masks, keep-side correction, and layer sets unchanged. For each approach, we select the bias with the highest validation identity similarity satisfying $\Delta\text{LPIPS}_{\text{ne}} \leq 0.005$ (table 9). The selected biases are 1.5 for JoyAI-Video-Edit, 2.5 for LoomVideo, and 2.0 for both FLUX.2 variants. Their test-set results are reported in table 4.

Approach	$\pi_{\mathcal{R}}$	$\bar{\rho}_{\text{edit}}$	u_{edit}	$\bar{\rho}_{\text{keep}}$	u_{keep}	B	$\bar{b}_{\text{edit}}$
JoyAI-Video-Edit	n/a	.0615	—	.0555	—	2.30	2.96
LoomVideo	.031	.0069	0.22	.0040	0.13	2.32	5.91
FLUX.2-klein	.286	.0988	0.35	.0290	0.10	4.48	2.68
FLUX.2-klein-base	.286	.0635	0.22	.0433	0.15	1.88	3.24
VACE	.033	.0590	1.79	.0228	0.69	5.99	3.21
Qwen-Image-Edit	.332	.0990	0.30	.1157	0.35	1.11	2.47
OmniGen2	.327	.0783	0.24	.0916	0.28	1.04	2.70

Table 7: **Full reference-attention diagnostics.** Measurements from seven base models on the head-swap validation split. B is the peak edit-to-keep reference-mass ratio (eq. (6)); $\bar{b}_{\text{edit}}$ is the mean log-inverse edit mass (eq. (7)). JoyAI-Video-Edit's uniform baseline and normalized masses are omitted because its key composition varies during streaming.

γ_{edit}	γ_{keep}=0 ID↑	γ_{keep}=0 LPIPS_{ne}↓	γ_{keep}=0.4 ID	γ_{keep}=0.4 LPIPS_{ne}	γ_{keep}=0.8 ID	γ_{keep}=0.8 LPIPS_{ne}
JoyAI-Video-Edit (B=2.3); base ID .557, LPIPS_{ne} .0663						
all layers ⋆						
0.4	.689	.0615	.683	.0625	.682	.0618
0.6	.705	.0635	.705	.0644	**.706**	**.0641**
0.8	.548	.0913$^{\times}$	.543	.0914$^{\times}$	.549	.0917$^{\times}$
second half						
0.4	.652	.0639	.651	.0638	.652	.0644
0.6	.675	.0678	.675	.0678	.674	.0674
0.8	.675	.0784$^{\times}$	.675	.0780$^{\times}$	.676	.0788$^{\times}$
LoomVideo (B=2.3, $\bar{b}$=5.91); base ID .653, LPIPS_{ne} .0700						
second half ⋆						
0.4	.742	.0696	.739	.0681	.739	.0666
0.6	.768	.0781$^{\times}$	.767	.0746	**.767**	**.0727**
0.8	.771	.1317$^{\times}$	.769	.1281$^{\times}$	.769	.1262$^{\times}$
all layers						
0.4	.766	.1306$^{\times}$	.768	.1250$^{\times}$	.789	.1139$^{\times}$
0.6	.824	.3187$^{\times}$	.825	.3094$^{\times}$	.826	.2781$^{\times}$
0.8	.847	.3334$^{\times}$	.847	.3150$^{\times}$	.847	.2854$^{\times}$
FLUX.2-klein (B=4.48); base ID .807, LPIPS_{ne} .0435						
second half ⋆						
0.4	.882	.0497$^{\times}$	.882	.0395	.882	.0387
0.6	.885	.0524$^{\times}$	.885	.0416	**.885**	**.0406**
0.8	.889	.0561$^{\times}$	.889	.0438	.889	.0428
all layers						
0.4	.888	.1079$^{\times}$	.892	.0484	.892	.0445
0.6	.888	.1779$^{\times}$	.893	.0644$^{\times}$	.893	.0545$^{\times}$
0.8	.890	.1918$^{\times}$	.891	.0780$^{\times}$	.891	.0686$^{\times}$
FLUX.2-klein-base (B=1.88); base ID .590, LPIPS_{ne} .1206						
all layers ⋆						
0.4	.797	.1900$^{\times}$	.866	.0687	.864	.0654
0.6	.879	.2115$^{\times}$	.877	.0764	**.878**	**.0674**
0.8	.873	.2273$^{\times}$	.877	.0979	.875	.0835
second half						
0.4	.704	.1373$^{\times}$	.773	.0941	.773	.0914
0.6	.791	.1439$^{\times}$	.812	.0958	.826	.0932
0.8	.824	.1434$^{\times}$	.844	.0996	.842	.0957
Held out from coefficient selection: $(\gamma_{\text{edit}}, \gamma_{\text{keep}}) = (0.6, 0.8)$						
VACE (6.0; half)	ID 0.484 → **0.810**		LPIPS_{ne} 0.045 → 0.044 ✓			
Qwen-Image-Edit (1.11; all)	ID 0.455 → **0.787**		LPIPS_{ne} 0.296 → 0.167 ✓			
OmniGen2 (1.04; all)	ID 0.413 → **0.841**		LPIPS_{ne} 0.237 → 0.124 ✓			

Table 8: **Joint calibration of correction coefficients and layer coverage.** Head-swap validation results. ⋆ marks rule-selected layers; bold marks the chosen configuration. LPIPS_{ne} denotes non-edit LPIPS, and $^{\times}$ marks an increase over the base exceeding 0.005.

D Task protocols and additional results

D.1 Head swapping

Protocol and metrics. HeadSwapBench (Wang et al., 2025) contains 1040 test clips and a disjoint validation split of 40 clips. Each clip contains 121 frames at 512×512 resolution. Inputs comprise a driving video and a reference head image; VACE receives a head-masked driving video. Image approaches are edited and evaluated on the middle frame.

Because head swapping replaces the entire head, including the hair, whereas AdaFace focuses only on the face, we additionally report whole-head similarity (Wang et al., 2026), which measures the cosine similarity between head-appearance embeddings of the generated result and the ground truth. Keypoint error uses 478 landmarks and is normalized by inter-ocular distance.

Task-fine-tuned model: DirectSwap. While the main paper evaluates general-purpose editors, we further apply RefGAP to a task-specific fine-tuned model. We evaluate DirectSwap (Wang et al., 2025) in its released inference pipeline, including its attention-based background compositing. For this mask-free model, we define edit queries from the reference-attention mass of the first denoising

b	JoyAI-Video-Edit all ID↑	JoyAI-Video-Edit all LPIPS_{ne}↓	LoomVideo second half ID↑	LoomVideo second half LPIPS_{ne}↓	FLUX.2-klein second half ID↑	FLUX.2-klein second half LPIPS_{ne}↓	FLUX.2-klein-base all ID↑	FLUX.2-klein-base all LPIPS_{ne}↓
base	.557	.0663	.653	.0700	.807	.0435	.590	.1206
0.5	.647	.0641	.661	.0681	.851	.0377	.806	.0667
1.0	.683	.0625	.689	.0676	.883	.0394	.861	.0659
1.5	**.703**	**.0616**	.717	.0669	.888	.0410	.880	.0667
2.0	.625	.0761$^{\times}$	.744	.0670	**.889**	**.0432**	**.881**	**.0698**
2.5	—	—	**.761**	**.0697**	.888	.0474	.871	.0763
3.0	—	—	.774	.0790$^{\times}$	.888	.0544$^{\times}$	.873	.0898

Table 9: **Fixed edit-side sweeps that select each approach's constant** (head-swap validation, 40 clips). $^{\times}$ marks $\Delta\text{LPIPS}_{\text{ne}} > 0.005$ relative to the base; bold marks the best admissible constant, the highest-identity value that satisfies the criterion, which is the constant used in table 4.

	ID (AdaFace)↑	Kpt.↓	LPIPS↓
DirectSwap	$0.7019 \rightarrow 0.7450$	$0.0289 \rightarrow 0.0371$	$0.1333 \rightarrow 0.1339$

Table 10: **DirectSwap on the HeadSwapBench test split.** Scores are reported as base → RefGAP using clip-level averaging. ID and Kpt. denote identity similarity and keypoint error, respectively. LPIPS is measured over the whole frame.

step, averaged over all attention heads: in each of layers 8–26 (zero-based), a query casts a vote if its mass lies above the 0.7 quantile of its latent frame, and it is an edit query if more than 20% of these layers vote for it. These are the defaults of DirectSwap's released background-compositing mask, so RefGAP adds no threshold of its own. The first step runs unmodified; the region is then fixed, and RefGAP acts on layers 15–29 for the remaining 29 of 30 sampling steps.

As shown in table 10, RefGAP increases identity similarity from 0.7019 to 0.7450. These gains accompany increased keypoint error, while whole-frame LPIPS changes little. The results show that reference-attention correction can also improve identity fidelity in a task-fine-tuned model, with remaining attribute-preservation trade-offs.

Whole-head similarity. Table 11 complements the identity similarity results in table 2. Whole-head similarity increases on all seven general-purpose approaches.

	JoyAI-Video-Edit	LoomVideo	FLUX.2-klein	FLUX.2-klein-base	VACE	Qwen-Image-Edit	OmniGen2	DirectSwap
base	0.809	0.816	0.861	0.757	0.786	0.734	0.581	0.894
RefGAP	**0.876**	**0.861**	**0.887**	**0.875**	**0.888**	**0.802**	**0.851**	0.894
win %	73	73	61	76	96	63	96	49

Table 11: **Whole-head similarity on HeadSwapBench.** Win % is the percentage of paired valid clips on which RefGAP outperforms the base. DirectSwap is included as a task-fine-tuned reference.

D.2 Face swapping

Inputs and preprocessing. We use the c23-compressed real videos from FaceForensics++ (Rossler et al., 2019), following the 1000 official target–source pairings of its FaceSwap subset. Both inputs are real videos; manipulated videos are used only to identify the official pairings. VACE receives a face-masked driving video.

Each target clip contains the first 33 frames. We apply a fixed face-centered square crop with side length twice the median face height across the clip and resize it to 512×512. The edit mask undergoes the same transformation. Image approaches edit the middle frame.

We select a reference frame with a near-neutral expression from the first 300 source frames, using MediaPipe blendshape scores to penalize facial expressions, particularly jaw opening and blinking. If no face is detected (21 pairs), we use the first frame.

Why face-centered crops. In the full 640×480 frames the face occupies a small part of the image, so at the resolutions the approaches process it spans few latent tokens: the edit has little room to render the source identity, and mouth and eye details degrade. The crop places the face at roughly a quarter of the image area, matching the head scale of the head-swap benchmark. We compared the two inputs on JoyAI-Video-Edit with the same neutral references (table 12). On these pairs the crop raises retrieval and identity similarity for both arms and lowers keypoint. All face-swap results in the paper use the crops.

input	ID-Retr. (%) ↑		ID-sim ↑		Kpt. ↓	
	base	RefGAP	base	RefGAP	base	RefGAP
full frame (640×480)	25.0	50.0	.109	.240	.085	.108
face-centered crop (512^2)	40.0	**80.0**	.248	**.434**	.058	.074

Table 12: **Full-frame versus face-centered inputs for face swapping**

D.3 Virtual try-on

Datasets and inputs. Our main evaluation uses 90 ViViD videos (Fang et al., 2024). Each clip contains 65 frames at 624×832 resolution (640×832 for LoomVideo). We additionally evaluate image approaches on 100 VITON-HD images (Choi et al., 2021).

Both evaluations use an unpaired setting: reference garments are permuted within category, and the model is instructed to replace the input garment with the reference garment. Models receive the unmasked input, except VACE, which uses its masked-input interface. For VITON-HD, inputs to FLUX.2-klein and FLUX.2-klein-base are padded to a square canvas, and outputs are cropped back before evaluation.

Additional evaluation on VITON-HD. We further evaluate the four image approaches on 100 VITON-HD images using the same coefficients and layer sets. As shown in table 13, DINO improves on all four approaches by 0.009–0.046, with LPIPS increases of at most 0.0056. These results support transfer to an additional image dataset without retuning.

Approach	DINO ↑	LPIPS ↓
FLUX.2-klein	0.382 → 0.391	0.0302 → 0.0318
FLUX.2-klein-base	0.407 → 0.426	0.0415 → 0.0417
Qwen-Image-Edit	0.216 → 0.225	0.0158 → 0.0160
OmniGen2	0.397 → 0.443	0.1339 → 0.1395

Table 13: **Virtual try-on on** 100 **VITON-HD images.** Scores are shown as base → RefGAP.

D.4 Background replacement

Inputs and preprocessing. We use 150 ground-truth videos from 58 subjects in the HeadSwap-Bench test split, selecting static-camera clips containing a single person. Each clip is paired with a person-free scene reference from a set of 30 SUN397 images (Xiao et al., 2010). SAM3 (Carion et al., 2025) person masks define the subject and background regions. Video approaches edit the original clips, while image approaches edit their middle frames. All approaches use the same clip-reference pairs.

E Uncertainty of the main quantitative results

We quantify uncertainty in the Base-to-Ours differences in table 2 using paired bootstrap intervals that account for shared identities and reference content. Metric definitions and task-specific protocols are provided in section D.

Paired estimates. Base and RefGAP use identical inputs, seeds, and masks. For each metric, we compute the clip-level difference $d_i = m_i^{\text{Ours}} - m_i^{\text{Base}}$ and average over the eligible paired observations. Frames within a video are aggregated before resampling; image approaches are evaluated on the middle frame. Repl. and Retr. differences are expressed in percentage points (pp). Estimates and intervals are computed from unrounded per-clip scores, rather than from the rounded values displayed in table 2.

Resampling units. We use task-specific units to represent dependencies between clips:

- **Head swap:** subjects, with 104 identities across 1,040 clips. OmniGen2's face-based metrics retain 103 identities.
- **Face swap:** a shared pool of target and source identities, accounting for identities that occur in both roles.
- **Virtual try-on:** input clips and reference garments, with garments assigned from other clips in the same category.
- **Background replacement:** subject identities and scene references, comprising 58 people and 30 distinct scene images.

Repl. is omitted for VACE in both identity tasks because its masked input does not retain the input identity.

Bootstrap procedure. Following the factor-wise resampling construction of the pigeonhole bootstrap (Owen, 2007), we resample the units of each factor with replacement. For crossed factors, the draws are independent, and observation i receives weight

$$w_i = \prod_f c_f(u_{i,f}),$$

where $c_f(u_{i,f})$ is the sampled multiplicity of its unit in factor f. Each replicate estimates

$$\widehat{\Delta}^* = \frac{\sum_i w_i d_i}{\sum_i w_i}.$$

For head swap, this reduces to a subject-level cluster bootstrap. For face swap, we draw multiplicities once from the shared identity pool and weight each pair by the product of its target and source multiplicities. We generate 10,000 replicates and report the 2.5th and 97.5th percentiles. These are pointwise intervals, not simultaneous intervals across all comparisons. We also compute an ordinary paired clip bootstrap for comparison.

Overall uncertainty. Table 14 reports 89 paired estimates and intervals. Relative to clip-level resampling, the dependence-aware procedure increases interval widths by a median factor of 1.85. The number of intervals excluding zero decreases from 76 to 62, indicating that treating clips as independent can overstate the precision of the estimated effects.

Identity fidelity. All 27 intervals for head-swap Repl. and ID and face-swap ID and Retr. exclude zero in the improving direction. This includes LoomVideo's face-swap ID gain of $+.017$ $[+.003, +.034]$. Face-swap Repl. improves with an interval excluding zero on all six approaches for which it is defined.

Preservation trade-offs. The intervals also identify preservation costs. Head-swap keypoint error increases on JoyAI-Video-Edit and both FLUX.2 variants. Face-swap keypoint error increases on these three approaches, Qwen-Image-Edit, and VACE; face-swap LPIPS increases on JoyAI-Video-Edit, both FLUX.2 variants, and VACE. Background LPIPS increases on JoyAI-Video-Edit, LoomVideo, and OmniGen2. Conversely, head-swap LPIPS improves on Qwen-Image-Edit, FLUX.2-klein-base, VACE, and OmniGen2, while OmniGen2 also improves face-swap and try-on LPIPS. Each change listed here has an interval excluding zero.

Task	Metric	JoyAI	LoomVideo	Klein	Klein-base	VACE	Qwen	OmniGen2
Head swap	Repl. (pp)↑	+30.9 [+21.5, +40.1]	+12.2 [+6.6, +17.6]	+8.6 [+5.1, +12.3]	+22.8 [+17.5, +27.4]	—	+16.1 [+10.1, +24.1]	+65.1 [+56.7, +71.4]
	ID↑	+.137 [+.111, +.169]	+.058 [+.045, +.069]	+.064 [+.049, +.088]	+.211 [+.176, +.247]	+.323 [+.290, +.362]	+.114 [+.071, +.167]	+.486 [+.449, +.525]
	Kpt.↓	+.042 [+.027, +.059]	−.009 [−.025, +.002]	+.055 [+.046, +.065]	+.068 [+.048, +.096]	+.001 [−.009, +.012]	+.000 [−.017, +.015]	+.002 [−.019, +.021]
	LPIPS↓	−.003 [−.010, +.006]	−.005 [−.013, +.000]	−.004 [−.011, +.005]	−.066 [−.093, −.037]	−.013 [−.016, −.011]	−.212 [−.246, −.171]	−.099 [−.117, −.079]
Face swap	Repl. (pp)↑	+36.8 [+31.4, +42.6]	+13.0 [+9.4, +17.1]	+2.2 [+0.8, +4.1]	+5.1 [+2.9, +7.8]	—	+16.7 [+11.7, +21.8]	+20.7 [+16.7, +24.9]
	ID↑	+.310 [+.293, +.327]	+.017 [+.003, +.034]	+.116 [+.102, +.131]	+.274 [+.244, +.305]	+.277 [+.262, +.294]	+.169 [+.137, +.202]	+.498 [+.474, +.521]
	Retr. (pp)↑	+49.1 [+43.6, +54.7]	+11.3 [+7.6, +15.3]	+6.1 [+3.6, +9.1]	+22.4 [+17.3, +27.6]	+11.7 [+8.3, +15.4]	+20.1 [+15.0, +25.3]	+71.6 [+66.5, +76.4]
	Kpt.↓	+.048 [+.045, +.052]	+.003 [−.002, +.008]	+.017 [+.015, +.019]	+.093 [+.081, +.106]	+.006 [+.001, +.012]	+.022 [+.013, +.030]	+.011 [−.001, +.023]
	LPIPS↓	+.022 [+.019, +.026]	−.001 [−.006, +.003]	+.013 [+.011, +.015]	+.006 [+.001, +.012]	+.001 [+.001, +.001]	−.004 [−.011, +.001]	−.044 [−.057, −.031]
Virtual try-on	DINO↑	+.016 [−.006, +.039]	+.010 [−.001, +.026]	+.009 [−.001, +.021]	+.056 [+.032, +.090]	+.055 [+.024, +.092]	+.035 [+.003, +.069]	+.078 [+.024, +.136]
	LPIPS↓	+.001 [−.009, +.009]	+.001 [−.006, +.004]	−.001 [−.006, +.002]	+.010 [−.001, +.023]	−.000 [−.000, +.000]	+.017 [−.002, +.043]	−.017 [−.034, −.003]
Background	DINO↑	+.045 [+.014, +.086]	+.123 [+.064, +.187]	+.001 [−.008, +.013]	+.019 [−.017, +.048]	+.006 [−.005, +.020]	−.257 [−.374, −.153]	+.270 [+.151, +.346]
	LPIPS↓	+.002 [+.001, +.003]	+.007 [+.004, +.011]	−.002 [−.004, +.000]	−.026 [−.050, +.001]	+.000 [−.000, +.000]	−.004 [−.021, +.011]	+.036 [+.006, +.068]

Table 14: **Paired differences with pointwise 95% bootstrap confidence intervals.** Cells show Ours−Base and the corresponding interval from 10,000 task-specific resamples. Green/red indicate intervals excluding zero in the improving/degrading direction; gray indicates intervals containing zero. Colors are determined before rounding. Repl. and Retr. differences are in percentage points. VACE Repl. is undefined and omitted. Sample counts vary by metric and detection availability.

Transfer beyond identity editing. Try-on DINO gains have intervals excluding zero on four approaches, while the intervals for JoyAI-Video-Edit, FLUX.2-klein, and LoomVideo include zero. For background replacement, DINO improves on JoyAI-Video-Edit, LoomVideo, and OmniGen2 with intervals excluding zero. The decrease on Qwen-Image-Edit also excludes zero ($-.257$ $[-.374, -.153]$), whereas the intervals for both FLUX.2 variants and VACE include zero. Thus, the transfer results support gains on several approaches, rather than uniform improvements across all settings.

The 14 comparisons whose intervals exclude zero only under clip-level resampling concern preservation or transfer metrics. For these comparisons, we report the estimated direction without claiming a reliably established improvement or degradation.

F Human Evaluation

Sampling details. For each approach and task, we first form a candidate pool of samples showing large visible differences between the two methods, with method labels hidden, and then randomly select items from this pool using a fixed seed while rotating over subject identities. We evaluate five items per approach for head swapping and ten for each other task, except that FLUX.2-klein contributes only two items each for try-on and background replacement because the two methods produce similar outputs on most samples. VACE background replacement is excluded because neither method performs the edit.

Unscreened audit. To measure what the screening leaves out, we repeated the protocol on a form built without it: two items per approach and task drawn at random from each evaluation set (56 items, VACE background replacement included), plus two attention checks and two A/B-swapped repeats, answered by five raters. All five passed both checks, agreed with one another on 95–99% of answers, and repeated their own judgment on 87% of the swapped repeats. For reference fidelity, 27.9% of judgments favored RefGAP, 3.6% favored the base, and 68.6% were ties, giving RefGAP

	RefGAP	base	both succeed	both fail	tie %
Reference fidelity	78 (27.9%)	10 (3.6%)	144	48	68.6
Pose / expression	20 (7.1%)	10 (3.6%)	178	72	89.3
Visual quality	0 (0.0%)	22 (7.9%)	246	12	92.1

Table 15: **Unscreened audit form.** Judgment counts from five raters on 56 randomly sampled items (two per approach and task), with no screening for visible differences.

an 88.6% share among decided judgments. For visual quality, 92.1% were ties, while all 22 decided judgments favored the base. Full results are reported in table 15.

The 22 base-favoring quality judgments are concentrated in five items, rather than 22 distinct examples. This is consistent with the base-favoring quality preference observed for LoomVideo face swapping in the screened evaluation (1/33/18, RefGAP/tie/base). However, table 2 shows mixed changes for LoomVideo face swapping: identity similarity improves, keypoint error increases slightly, and non-face LPIPS decreases slightly. These metrics therefore do not fully capture the perceived quality differences.

Quality controls. Each form includes two attention checks with an unedited input as one candidate. One of 55 raters fails a check and is excluded. Three additional identical-output controls receive tie responses in all 45 judgments.

Additional statistics. Table 17 reports item-level preference rates with 95% confidence intervals from $10{,}000$ bootstrap resamples of items. Preference rates exclude ties separately for each question; items with only tied responses to a question are omitted from its preference estimate. The reference-fidelity tie rate is reported separately.

Table 16 separates the two responses pooled as ties in table 3. Pose/expression receives 46 both-fail judgments on head swapping and 138 on face swapping, showing why ties should not be interpreted as successful preservation.

Task	Question	RefGAP	both succeed	both fail	base	total
Head swap	closer to reference	157	23	1	3	184
	pose / expression	27	109	46	3	185
	visual quality	37	131	7	10	185
Face swap	closer to reference	278	89	7	8	382
	pose / expression	19	185	138	40	382
	visual quality	32	312	8	30	382
Try-on	closer to reference	249	37	13	40	339
	pose / expression	9	274	53	3	339
	visual quality	24	298	11	6	339
Background	closer to reference	127	101	20	29	277
	pose / expression	5	217	52	3	277
	visual quality	17	242	8	9	276

Table 16: **Breakdown of tied judgments.** The "both succeed" and "both fail" responses are reported separately and sum to the tie counts in table 3. Entries report the number of rater responses.

Face-swap subgroup analysis. We further group face-swap items by whether the base output is closer to the reference identity or the input identity, excluding two items without a detected face in the base. Reference-fidelity judgments favor RefGAP in both groups. For pose/expression, the base receives 25 preferences versus 3 for RefGAP when its output is closer to the input identity. When it is closer to the reference identity, preferences are nearly balanced (16 for RefGAP versus 15 for the base, with 234 ties). This analysis provides the breakdown underlying the observation in section 4.4.

Task	Ref.	Pose/expr.	Quality	Ties
Head swap	93.5 [84.7,100]	86.0 [66.0,100]	86.2 [71.9,97.9]	12.6 [3.6,23.4]
Face swap	93.6 [87.2,98.7]	39.7 [23.0,57.0]	56.1 [40.0,72.2]	26.0 [16.9,35.4]
Try-on	85.2 [75.4,93.8]	71.4 [42.9,100]	85.0 [70.0,100]	14.6 [6.8,23.5]
Background	82.2 [68.8,93.8]	40.0 [0.0,80.0]	52.2 [30.0,74.4]	45.7 [32.7,58.6]

Table 17: **Item-level human preferences pooled across approaches (%).** Ref., Pose/expr., and Quality report preference for RefGAP after excluding ties. Ties reports the reference-fidelity tie rate. Brackets show 95% item-bootstrap confidence intervals.

G ATTRIBUTE TRADE-OFFS AND FAILURE ANALYSIS

In head swapping, keypoint degradation is concentrated in JoyAI-Video-Edit, FLUX.2-klein, and FLUX.2-klein-base, whereas face-swap keypoint error increases on most approaches. This difference partly reflects the face-swap evaluation protocol, in which the output is compared with the input geometry while being encouraged to adopt the reference identity. We first disentangle this identity-geometry confound and show that the remaining keypoint and expression costs are concentrated in the same three approaches as in head swapping. We then examine whether delaying the attention correction reduces these costs. Finally, we analyze the background-replacement failure of Qwen-Image-Edit and the occasional blurred-face failures of OmniGen2.

G.1 IDENTITY-ATTRIBUTE TRADE-OFFS IN FACE SWAPPING

FF++ does not provide ground-truth edited frames, so face-swap keypoints are evaluated against the driving input rather than the reference identity. Although the landmarks are scale-normalized, they remain sensitive to identity-specific face shape. Consequently, successful identity transfer can increase keypoint error even when the input pose is preserved.

To quantify this effect, table 18 relates the clip-level change in keypoint error to the corresponding identity gain. For Qwen-Image-Edit, LoomVideo, and OmniGen2, the increase in keypoint error is concentrated among clips with large identity gains: their lowest-gain quartiles have near-zero or negative changes, and only small costs remain when $|\Delta\text{ID}| < 0.05$. These patterns are consistent with identity-related geometry changes contributing to the measured keypoint increase.

JoyAI-Video-Edit and FLUX.2-klein-base retain keypoint increases of 0.029 and 0.052, respectively, when identity similarity changes by less than 0.05. Together with the smaller residual increase on FLUX.2-klein, this matches the approach-level pattern observed in head swapping, where a ground-truth target removes the identity confound. After accounting for identity gain, keypoint degradation is likewise concentrated in these three approaches. VACE is listed for completeness but not discussed, since its masked input provides no expression signal.

Approach	Δkpt	ΔID	lowest ΔID quartile	highest ΔID quartile	$\lvert\Delta\text{ID}\rvert < 0.05$ (n)
JoyAI-Video-Edit	$+0.048$	$+0.310$	$+0.037$	$+0.064$	$+0.029$ (44)
LoomVideo	$+0.003$	$+0.017$	-0.010	$+0.012$	$+0.006$ (350)
FLUX.2-klein	$+0.017$	$+0.116$	$+0.011$	$+0.026$	$+0.012$ (348)
FLUX.2-klein-base	$+0.094$	$+0.274$	$+0.055$	$+0.111$	$+0.052$ (230)
VACE	$+0.006$	$+0.278$	$+0.013$	$+0.002$	$+0.025$ (13)
Qwen-Image-Edit	$+0.021$	$+0.169$	-0.020	$+0.068$	$+0.012$ (338)
OmniGen2	$+0.011$	$+0.498$	-0.008	$+0.022$	$+0.005$ (16)

Table 18: **Face-swap keypoint change versus identity gain.** Per clip, Δkpt is RefGAP minus base keypoint error against the input subject (lower is better), and ΔID is RefGAP minus base AdaFace similarity to the reference image given to the model. The quartile columns report the average Δkpt within the lowest and highest quarters of ΔID. The last column reports the average over clips whose identity similarity changes by less than 0.05, with the corresponding count in parentheses.

HeadSwapBench	ID↑	kpt↓	LPIPS↓
FLUX.2-klein (4-step schedule)			
base	0.726	**0.097**	0.222
RefGAP (γ_{edit}=0.6, all steps)	**0.790**	0.153	**0.217**
RefGAP, skip step 1	0.753	0.101	0.221
JoyAI-Video-Edit (2-step schedule)			
base	0.507	**0.043**	0.218
RefGAP (γ_{edit}=0.6, all steps)	**0.644**	0.085	**0.215**
RefGAP, skip step 1	0.613	0.052	0.217
FLUX.2-klein-base (50-step schedule)			
base	0.588	**0.169**	0.326
RefGAP (γ_{edit}=0.6, all steps)	**0.799**	0.236	**0.261**
RefGAP, skip step 1	0.798	0.235	0.264
RefGAP, skip first $10/50$ steps	0.735	0.183	0.320

Table 19: **Disabling correction at early denoising steps.** Results on 1040 head-swap test clips. "Skip" disables the attention correction at the indicated steps; denoising still proceeds normally. LPIPS is measured over the whole frame. Bold marks the best value within each approach.

G.2 Denoising-step interventions

We next test whether the geometric cost can be reduced by delaying the attention correction. The full denoising trajectory is preserved; the correction is disabled only during selected initial steps.

As shown in table 19, skipping the first step reduces keypoint error from 0.085 to 0.052 on JoyAI-Video-Edit and from 0.153 to 0.101 on FLUX.2-klein. These improvements are accompanied by reductions in identity similarity from 0.644 to 0.613 and from 0.790 to 0.753, respectively.

FLUX.2-klein-base uses a longer, 50-step schedule, for which skipping only the first step has little effect. Skipping the first ten steps reduces keypoint error from 0.236 to 0.183, but also lowers identity similarity from 0.799 to 0.735.

Thus, delaying the correction can partially preserve input geometry, but at the cost of weaker identity transfer. Because this trade-off depends on the approach and sampling schedule, we retain correction at all steps to prioritize recognizable reference identity.

G.3 Background replacement failure

Background replacement differs from head swapping, face swapping, and virtual try-on: the reference specifies the surrounding scene rather than the person or their appearance, while the person must be preserved. This shift in the role of the reference may affect how an approach responds to stronger reference conditioning.

On Qwen-Image-Edit, the shared setting reduces DINO from 0.299 to 0.042. Restricting correction to the second half does not recover fidelity on the diagnostic clips. A fixed edit-side bias of $b = 2.0$ also produces a similar degradation, indicating that the failure is not specific to adaptive edit-side strength (See table 20).

			diagnostic clips ($n = 30$)		all clips ($n = 150$)
setting	layers	edit-side strength	DINO ↑	subject LPIPS ↓	DINO ↑
base	—	—	.473	.0384	.299
RefGAP (shared setting)	all	online, $\gamma_{\text{edit}} = 0.6$	.112	.0242	.042
RefGAP, second half	30–59	online, $\gamma_{\text{edit}} = 0.6$	.106	.0127	—
fixed edit-side bias	all	constant, $b = 2.0$	.040	.0174	.037

Table 20: **Qwen-Image-Edit background replacement under alternative settings.** The diagnostic clips are the first 30 of the 150-clip list, not selected on outcome.

Other approaches improve on this task, so the task difference alone does not explain the failure. The results suggest an approach-dependent limitation when transferring the correction from person-related edits to scene replacement; its underlying cause remains unresolved.

G.4 Rare head-swap collapses

OmniGen2 occasionally produces missing or heavily smeared faces in both the base and RefGAP settings. Screening followed by manual inspection identifies $3/1040$ such outputs for the base and $8/1040$ for RefGAP, indicating an existing failure mode that becomes more frequent under the correction. As shown in fig. 4, changing the sampling seed can recover plausible faces, while repeating the same seed reproduces the failure.

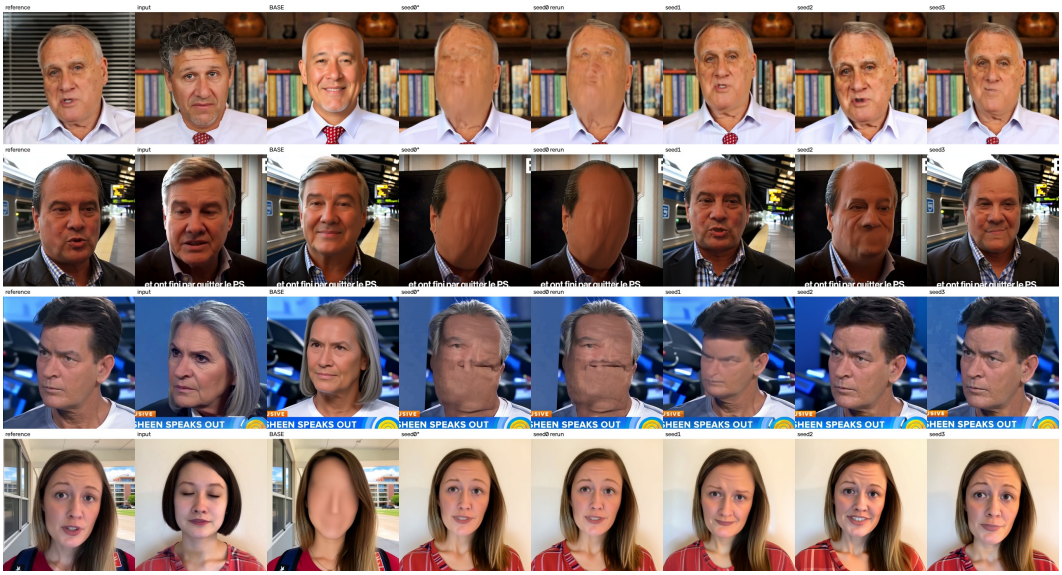


Figure 4: **Seed sensitivity of OmniGen2 head-swap outputs.** Columns show the reference, input, base, RefGAP at seed 0, a repeat at the same seed, and three alternative seeds. Repeated seeds reproduce the failures, while alternative seeds can recover plausible faces.

H Qualitative results

Figure 5 extends the examples in fig. 3 to all seven approaches. Figure 6 provides an additional set of examples across the same four tasks.

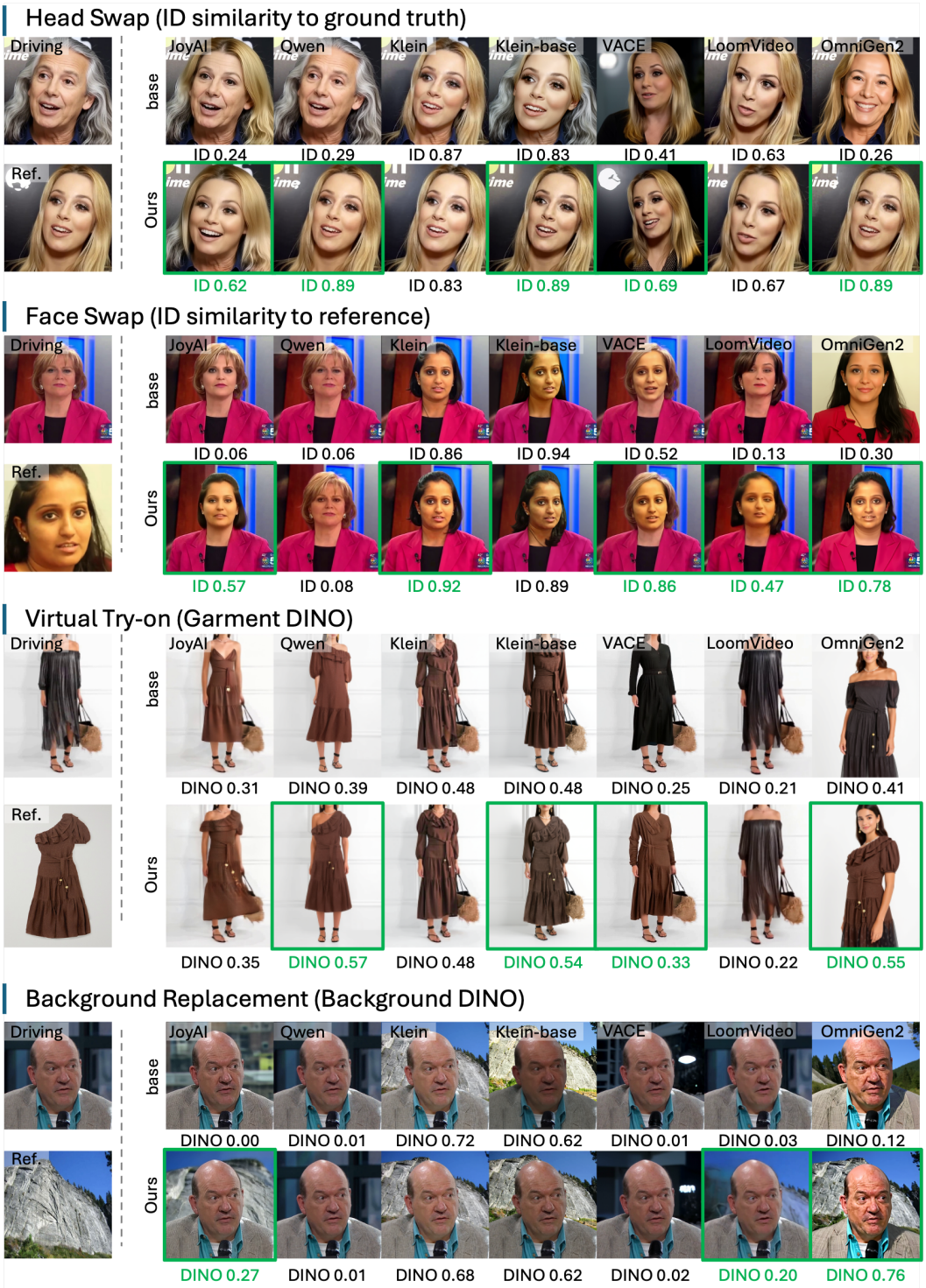


Figure 5: **Complete seven-approach comparison for the main-paper examples.** Each block shows the input, reference, and base/RefGAP outputs, with reference-fidelity scores beneath. Green borders indicate gains ≥ 0.05.



Figure 6: **Additional qualitative examples across seven approaches and four tasks.** The layout and annotations follow fig. 5.