REFFLOW: TRAINING-FREE REFERENCE-GUIDED IMAGE EDITING VIA RESIDUAL INSTANCE TRANSPORT

Anonymous authors
Paper under double-blind review

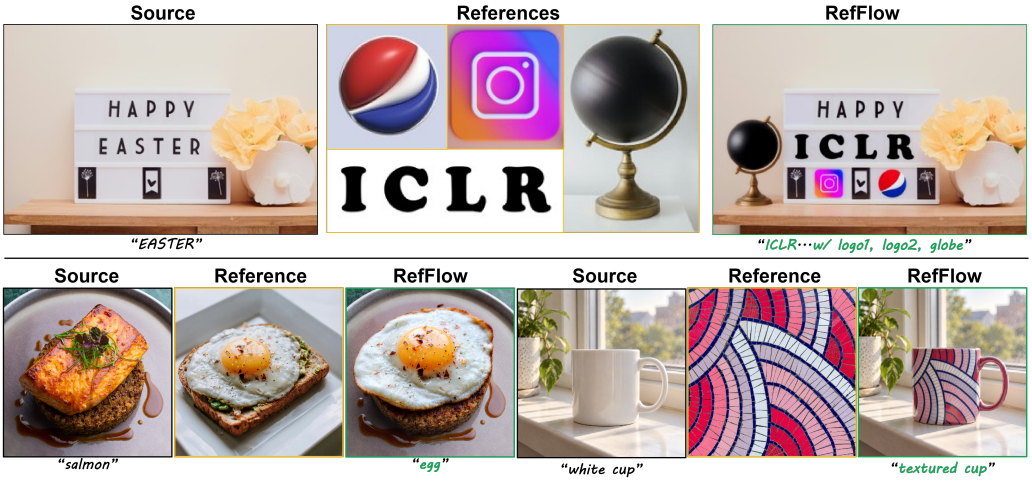

Figure 1: **Reference-guided editing with REFFLOW.** Our training-free framework supports a broad range of image editing tasks guided by single or multiple visual references.

ABSTRACT

Recent training-free image editing methods with pretrained generative models enable flexible text-driven manipulation, while text guidance alone cannot reproduce the appearance of a specific visual instance. Moreover, even with a visual reference, it remains challenging to simultaneously achieve task diversity and reference fidelity. We introduce REFFLOW, a training-free framework for reference-guided image editing via residual instance transport. REFFLOW augments semantic editing with a reference-induced velocity residual, implemented through Relationally Consistent Reference Value Transport that adaptively transfers reference appearance via target-adaptive attention retrieval and relational refinement. This design preserves the target editing trajectory while enabling instance-specific visual control without any additional training. REFFLOW is compatible with various pretrained models and supports diverse tasks, including object replacement, typography editing, texture transfer, and multi-reference composition. Experiments on RefEdit-Eval and DreamBooth demonstrate improved reference fidelity, editing success, and source preservation over existing training-free and trained baselines.

1 INTRODUCTION

Recent studies have demonstrated that pretrained text-to-image (T2I) models provide a powerful generative prior for training-free image editing (Kulikov et al., 2025; Kim et al., 2026). However, text prompts primarily specify high-level semantic intent rather than instance-specific visual details. Fine-grained attributes (such as customized typography styles, specific surface textures, or exact object identities) remain inherently underdetermined when guided solely by language. Visual references resolve this ambiguity by offering spatial and stylistic evidence beyond textual descriptions.

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.

Existing training-free methods incorporate visual references through diverse interfaces. SpecRef (Chen & Huang, 2023) applies masked attention control during T2I model diffusion denoising; VicoEdit (Song et al., 2026) and ReInversion (Li et al., 2025a) leverage visual context within pretrained image editing backbones (Black Forest Labs et al., 2025). In summary, existing T2I-based reference-guided methods generally introduce reference appearance by replacing attention outputs in the edited region with features retrieved through cross-image attention. However, without explicit coordination among correspondences, ambiguous matches can introduce inconsistent details when reference and target structures differ, making reference fidelity difficult to maintain across diverse editing tasks, as we further demonstrate in Figure 3.

To address these limitations, we introduce REFFLOW, a training-free framework that formulates reference-guided editing as *residual instance transport*. We preserve the text-guided semantic editing velocity as shown in equation 2 while injecting a spatially localized reference residual correction, see equation 3. Defined as the difference between target velocities evaluated with and without reference intervention at the same state, this residual isolates the reference contribution and enables explicit control over its strength and spatial support on a frozen T2I backbone. This design allows reference details to adapt to task-specific structures while the semantic component drives the requested transformation, supporting high-fidelity reference guidance across diverse editing tasks without inversion or training.

To construct this residual, we propose *Relationally Consistent Reference Value Transport*. Inspired by DreamMatcher (Nam et al., 2024), we decouple reference appearance retrieval from target feature aggregation, preserving the target's native attention structure. At each guided timestep, target queries establish soft correspondences that map reference values into target coordinates, allowing reference appearance to adapt to the evolving edit. However, independent matches may assign inconsistent reference details to related target regions (Vayer et al., 2019). We therefore jointly refine correspondences using within-image feature relations, encouraging compatible matches across parts—for example, keeping the head, torso, and limb features coherent when adapting a referenced dog to a different pose. This improves reference fidelity under diverse structural changes.

Furthermore, our formulation naturally supports multi-reference composition via spatially gated value fusion without increasing the target attention sequence length. Extensive evaluations demonstrate that REFFLOW generalizes beyond conventional single object placement editing to complex tasks such as typography editing and texture transfer, achieving superior reference fidelity and editing accuracy while maintaining background integrity. To the best of our knowledge, REFFLOW is the first method to achieve training-free, multi-reference image editing across diverse, complex scenarios using a completely frozen T2I model, see Figure 1.

Our main contributions are summarized as follows:

- We introduce a simple yet effective reference-induced velocity residual that incorporates explicit reference guidance, enabling training-free reference-guided editing applicable to different flow-based T2I backbones.
- We propose Relationally Consistent Reference Value Transport, which refines target-reference correspondence beyond independent token retrieval by preserving within-image feature relations, enabling adaptive and structurally consistent reference guidance.
- We demonstrate that REFFLOW generalizes beyond conventional object replacement to diverse reference-guided editing scenarios, including typography editing, texture transfer, and multi-reference composition. Extensive experiments show improved reference fidelity, editing success, and source preservation over existing methods.

2 RELATED WORK

2.1 TEXT-GUIDED IMAGE EDITING

Early text-guided editing relies on learned instruction-following models (Brooks et al., 2023). To eliminate training costs, training-free diffusion approaches manipulate editing trajectories via noise perturbation (SDEdit) (Meng et al., 2022), attention/feature injection (Prompt-to-Prompt, Plug-and-Play) (Hertz et al., 2023; Tumanyan et al., 2023), correspondence-aware noise correction (Alimohammadi et al., 2025), unconditional embedding optimization (Null-text inversion) (Mokady et al.,

2023), or key-value caching (KV-Edit) (Zhu et al., 2025). While effective at preserving background consistency, existing KV-based methods operate mainly by directly substituting cached keys and values, which constrains them to regions where the source content can be reused and limits their applicability to flexible appearance transfer. Recently, flow-based models have advanced training-free editing: FlowEdit (Kulikov et al., 2025) constructs inversion-free transport trajectories, while FlowAlign (Kim et al., 2026) introduces trajectory regularization. Subsequent works further refine flow dynamics via conditioned velocity correction (CVC) (Li et al., 2025b), fidelity–steering decomposition (FlowSlider) (Endo et al., 2026), low-energy transport (ChordEdit) (Lu et al., 2026), noise structure optimization (DNAEdit, SNR-Edit) (Xie et al., 2025; Jiang et al., 2026), and frequency-aware wavelet control (Tang et al., 2026). Despite their flexibility in driving high-level semantic editing, text-guided training-free methods rely solely on language prompts, which are inherently ambiguous and fundamentally incapable of executing fine-grained, instance-specific edits.

2.2 Reference-guided Image Editing

To supply precise visual evidence, learned reference-guided methods incorporate exemplar identity and detail features through dedicated conditioning modules (Chen et al., 2024; He et al., 2024) or joint image–text pretraining (Black Forest Labs et al., 2025; Xia et al., 2026; Qwen Team, 2025a). Among training-free alternatives, SpecRef (Chen & Huang, 2023) enforces masked attention control within inversion pipelines, ReInversion (Li et al., 2025a) performs alternating source/reference denoising, and VicoEdit (Song et al., 2026) integrates visual context with concept alignment. A parallel line of work establishes explicit correspondences for editing: MasaCtrl (Cao et al., 2023) converts self-attention into mutual self-attention to query correlated source contents, while Cora (Alimohammadi et al., 2025) uses semantic correspondence for noise correction in few-step editing. However, these methods either swap attention features or switch conditioning, without refining the reference-to-target attention correspondence itself. In contrast, our method treats the attention correspondence as an explicit per-step optimization variable, refined under within-image relational consistency and embedded into a residual velocity correction rather than replacing the semantic editing direction, thereby preserving the original edit intent while allowing reference appearance to follow the evolving target without additional training.

3 Preliminaries

FlowEdit (Kulikov et al., 2025) enables inversion-free text-guided editing with a pretrained flow model. Let $v_\theta(\boldsymbol{z}, t; c)$ denote the frozen velocity predictor conditioned on text c. Given a source latent $\boldsymbol{z}^{\mathrm{src}}$, FlowEdit initializes the editing state as $\boldsymbol{y}_1 = \boldsymbol{z}^{\mathrm{src}}$ and evolves it as t decreases from 1 to 0. At each timestep, it constructs two coupled inputs using $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$:

$$\boldsymbol{z}_t^{\mathrm{src}} = (1-t)\boldsymbol{z}^{\mathrm{src}} + t\boldsymbol{\epsilon}, \qquad \boldsymbol{z}_t^{\mathrm{tgt}} = \boldsymbol{y}_t + \boldsymbol{z}_t^{\mathrm{src}} - \boldsymbol{z}^{\mathrm{src}}. \tag{1}$$

This construction applies a common perturbation to the source and editing states.

With source and target text conditions c^{src} and c^{tgt}, FlowEdit updates the editing state through the difference between their predicted velocities:

$$\begin{aligned} \boldsymbol{y}_{t_{i-1}} &= \boldsymbol{y}_{t_i} + (t_{i-1} - t_i)\boldsymbol{u}_{t_i}^{\mathrm{FE}}, \\ \boldsymbol{u}_t^{\mathrm{FE}} &= \mathbb{E}_{\boldsymbol{\epsilon}}\left[v_\theta(\boldsymbol{z}_t^{\mathrm{tgt}}, t; c^{\mathrm{tgt}}) - v_\theta(\boldsymbol{z}_t^{\mathrm{src}}, t; c^{\mathrm{src}})\right]. \end{aligned} \tag{2}$$

The expectation is estimated by Monte Carlo sampling, and $\boldsymbol{y}_0$ is decoded into the edited image. This velocity difference defines the semantic editing component on which RefFlow builds.

4 RefFlow

FlowEdit (Kulikov et al., 2025) constructs a text-guided velocity field for semantic editing, yet leaves instance-specific visual details unspecified. Intuitively, incorporating reference guidance while preserving overall editing harmony requires only a localized residual force to gently nudge the trajectory toward the reference appearance. We thus formulate reference editing as a residual perturbation over the semantic trajectory, constructed by transporting reference values into target token coordinates. This spatially gated residual achieves fine-grained instance control without altering the underlying semantic objective, using a completely frozen flow model.

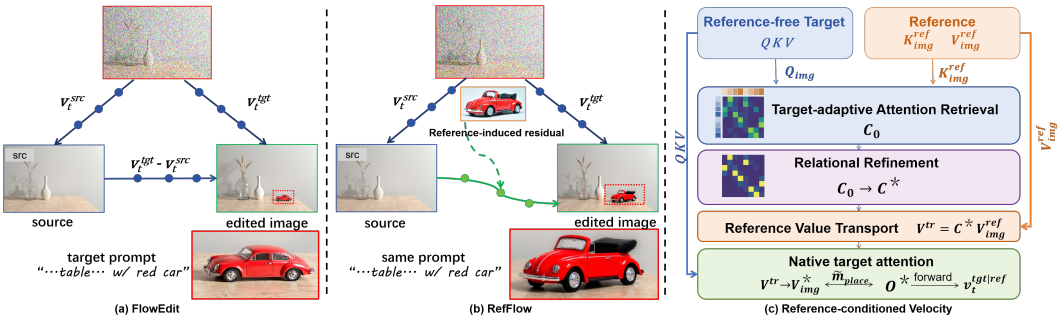

Figure 2: **Overview. (a)** FlowEdit (Kulikov et al., 2025) follows a semantic editing direction from the source image and target prompt, but does not control instance appearance. **(b)** REFFLOW adds a reference-induced residual under the same source and target prompt, steering the edit toward the reference instance. **(c)** The reference-conditioned target velocity $\boldsymbol{v}_t^{\mathrm{tgt|ref}}$ is constructed by target-adaptive attention retrieval, relational refinement, reference value transport, and native target attention.

4.1 REFERENCE-INDUCED RESIDUAL EDITING DYNAMICS

Let $\boldsymbol{v}_t^{\mathrm{src}} \equiv v_\theta(\boldsymbol{z}_t^{\mathrm{src}}, t; c^{\mathrm{src}})$ and $\boldsymbol{v}_t^{\mathrm{tgt}} \equiv v_\theta(\boldsymbol{z}_t^{\mathrm{tgt}}, t; c^{\mathrm{tgt}})$ denote the FlowEdit source and reference-free target velocities. We obtain $\boldsymbol{v}_t^{\mathrm{tgt|ref}}$ by rerunning the target branch at the same $\boldsymbol{z}_t^{\mathrm{tgt}}$ with the attention intervention in Section 4.2. Under otherwise identical conditions, their difference isolates the reference-induced velocity change:

$$\Delta \boldsymbol{v}_t = \underbrace{\boldsymbol{v}_t^{\mathrm{tgt}} - \boldsymbol{v}_t^{\mathrm{src}}}_{\text{semantic editing}} + \lambda \boldsymbol{m}_{\mathrm{place}} \odot \underbrace{\left(\boldsymbol{v}_t^{\mathrm{tgt|ref}} - \boldsymbol{v}_t^{\mathrm{tgt}}\right)}_{\text{reference-induced residual}}, \tag{3}$$

where λ controls reference strength and $\boldsymbol{m}_{\mathrm{place}}$ gates the residual spatially. Using the FlowEdit Monte Carlo estimator, we set $\boldsymbol{u}_t^{\mathrm{RF}} = \mathbb{E}_{\boldsymbol{\epsilon}}[\Delta \boldsymbol{v}_t]$ in equation 2. When $\lambda = 0$, REFFLOW reduces to FlowEdit. Figure 2 illustrates the formulation.

4.2 RELATIONALLY CONSISTENT REFERENCE VALUE TRANSPORT

To construct the reference-conditioned target velocity, we transport reference values into target-token coordinates using the current target attention structure. Here, $\boldsymbol{m}_{\mathrm{ref}}$ is the reference mask, $\boldsymbol{m}_{\mathrm{place}}$ is the target placement mask, and $\tilde{\boldsymbol{m}}_{\mathrm{place}}$ aligns $\boldsymbol{m}_{\mathrm{ref}}$ to $\boldsymbol{m}_{\mathrm{place}}$. Both masks can be externally provided or automatically estimated (Appendix A).

We first establish target-adaptive soft correspondences through attention interactions between target queries and reference keys. Since independent token-wise matches may be locally plausible but relationally inconsistent, we refine them by jointly preserving the initial feature matches and within-image relations. The refined correspondences are then used to transport reference values into the native target attention.

Target-adaptive Attention Retrieval At each reference-enabled timestep, the reference branch uses the same noise draw as the source–target pair:

$$\boldsymbol{z}_t^{\mathrm{ref}} = (1-t)\boldsymbol{z}^{\mathrm{ref}} + t\boldsymbol{\epsilon}. \tag{4}$$

Here, $\boldsymbol{z}^{\mathrm{ref}}$ denotes the spatially prepared reference latent. Its forward pass supplies the reference image keys $\boldsymbol{K}_{\mathrm{img}}^{\mathrm{ref}}$ and values $\boldsymbol{V}_{\mathrm{img}}^{\mathrm{ref}}$; the reference velocity itself does not enter the editing update. The reference-free target forward at $\boldsymbol{z}_t^{\mathrm{tgt}}$ supplies $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$. We omit timestep, layer, and head indices and retain the backbone's native query and key normalization.

Using pre-RoPE image features, we compute

$$S_{ij} = \frac{\alpha}{\sqrt{d}} \left\langle \boldsymbol{Q}_{\mathrm{img},i}, \boldsymbol{K}^{\mathrm{ref}}_{\mathrm{img},j} \right\rangle, \tag{5}$$

$$C^0_{ij} = \frac{\boldsymbol{m}_{\mathrm{ref}}(j) \exp(S_{ij})}{\sum_{j'} \boldsymbol{m}_{\mathrm{ref}}(j') \exp(S_{ij'})}. \tag{6}$$

Here, d is the head dimension, α scales the matching scores, and $\boldsymbol{m}_{\mathrm{ref}}$ restricts retrieval to valid reference-image tokens. The resulting correspondence adapts reference retrieval to the current target state, but its row-wise normalization treats target positions independently.

Relational correspondence refinement. We coordinate these matches through feature relations, following the principle of combining feature and structural matching (Vayer et al., 2019). Let $\widehat{\boldsymbol{K}}_{\mathrm{img}}$ and $\widehat{\boldsymbol{K}}^{\mathrm{ref}}_{\mathrm{img}}$ denote row-wise ℓ_2-normalized pre-RoPE image keys from the target and reference branches, respectively. We define

$$\boldsymbol{G}^{\mathrm{tgt}} = \widehat{\boldsymbol{K}}_{\mathrm{img}} \widehat{\boldsymbol{K}}^{\top}_{\mathrm{img}}, \qquad \boldsymbol{G}^{\mathrm{ref}} = \widehat{\boldsymbol{K}}^{\mathrm{ref}}_{\mathrm{img}} \left(\widehat{\boldsymbol{K}}^{\mathrm{ref}}_{\mathrm{img}} \right)^{\top}. \tag{7}$$

The mapped relation $\left(\boldsymbol{C}\boldsymbol{G}^{\mathrm{ref}}\boldsymbol{C}^{\top}\right)_{i\ell}$ measures the expected reference-key similarity under the correspondence distributions of target positions i and ℓ. Matching this quantity to $(\boldsymbol{G}^{\mathrm{tgt}})_{i\ell}$ encourages the retrieved reference content to preserve the corresponding target feature relation.

Let $\mathcal{F}$ denote the foreground target rows selected by the attention gate $\tilde{\boldsymbol{m}}_{\mathrm{place}}$. We compare their non-self relations to a fixed subset of foreground landmark tokens. For the resulting pair set $\mathcal{P}$, the spatial weights are

$$\begin{aligned} w_{i\ell} &= \mathbf{1}[(i,\ell) \in \mathcal{P}] \tilde{\boldsymbol{m}}_{\mathrm{place}}(i) \tilde{\boldsymbol{m}}_{\mathrm{place}}(\ell), \\ W_{i\ell} &= \sqrt{\frac{w_{i\ell}}{\sum_{u,v} w_{uv}}}. \end{aligned} \tag{8}$$

Thus, $\boldsymbol{W}$ provides normalized mask-based weights for the selected relations. We refine the correspondence through

$$\begin{aligned} \min_{\boldsymbol{C}} \quad & \frac{1}{|\mathcal{F}|} \sum_{i \in \mathcal{F}} D_{\mathrm{KL}} \left(\boldsymbol{C}_i \| \boldsymbol{C}^0_i \right) \\ & + \beta \left\| \boldsymbol{W} \odot \left(\boldsymbol{G}^{\mathrm{tgt}} - \boldsymbol{C}\boldsymbol{G}^{\mathrm{ref}}\boldsymbol{C}^{\top} \right) \right\|_F^2 \\ \text{s.t.} \quad & \boldsymbol{C} \geq 0, \qquad \boldsymbol{C}\mathbf{1} = \mathbf{1}, \\ & C_{ij} = 0 \quad \text{if } \boldsymbol{m}_{\mathrm{ref}}(j) = 0, \\ & \boldsymbol{C}_i = \boldsymbol{C}^0_i \quad \text{for } i \notin \mathcal{F}. \end{aligned} \tag{9}$$

The KL term preserves the initial feature matching, while the relation term couples correspondence decisions across target positions. For the selected relations, we evaluate

$$\left(\boldsymbol{C}\widehat{\boldsymbol{K}}^{\mathrm{ref}}_{\mathrm{img}} \right) \left(\boldsymbol{C}\widehat{\boldsymbol{K}}^{\mathrm{ref}}_{\mathrm{img}} \right)^{\top}$$

without renormalizing the mixed keys.

Starting from $\boldsymbol{C}^0$, a finite mirror-descent solve produces $\boldsymbol{C}^\star$. Only the correspondence is updated; cached features, the editing state, and model parameters remain fixed. Refinement is applied at selected intervention layers, while the remaining layers retain $\boldsymbol{C}^0$. Setting $\beta = 0$ recovers the initial correspondence. The refined correspondence transports reference image values as

$$\boldsymbol{V}^{\mathrm{tr}} = \boldsymbol{C}^\star \boldsymbol{V}^{\mathrm{ref}}_{\mathrm{img}}. \tag{10}$$

Spatially gated value intervention. $\boldsymbol{m}_{\mathrm{place}}$ gates the reference-induced velocity residual, while the aligned mask $\tilde{\boldsymbol{m}}_{\mathrm{place}}$ gates the attention intervention. Both are resampled to the corresponding feature grids and broadcast over channels.

We rerun the target network at the same target input $\boldsymbol{z}^{\mathrm{tgt}}_t$, timestep, text condition, and guidance settings. At each intervention layer, the target queries, keys, and text values are retained, while only

the image value slots are modified:

$$\begin{aligned} \boldsymbol{Q}^* = \boldsymbol{Q}, \qquad \boldsymbol{K}^* = \boldsymbol{K}, \qquad \boldsymbol{V}^*_{\text{txt}} = \boldsymbol{V}_{\text{txt}}, \\ \boldsymbol{V}^*_{\text{img}} = (1 - \tilde{\boldsymbol{m}}_{\text{place}}) \odot \boldsymbol{V}_{\text{img}} + \tilde{\boldsymbol{m}}_{\text{place}} \odot \boldsymbol{V}^{\text{tr}}. \end{aligned} \tag{11}$$

Let $\boldsymbol{V}^*$ denote the full value tensor obtained by combining $\boldsymbol{V}^*_{\text{txt}}$ and $\boldsymbol{V}^*_{\text{img}}$.

Since attention mixes values across positions, we additionally gate the induced correction at the receiving query positions:

$$\begin{aligned} \boldsymbol{A} &= \text{softmax}\left(\frac{\rho(\boldsymbol{Q})\rho(\boldsymbol{K})^\top}{\sqrt{d}}\right), \\ \boldsymbol{O}^* &= \boldsymbol{A}\boldsymbol{V} + \tilde{\boldsymbol{m}}_{\text{place}} \odot (\boldsymbol{A}\boldsymbol{V}^* - \boldsymbol{A}\boldsymbol{V}). \end{aligned} \tag{12}$$

Here, ρ applies the backbone's native RoPE, and the output gate is zero at text-token positions. The resulting target velocity is denoted by $\boldsymbol{v}_t^{\text{tgt}|\text{ref}}$ and enters equation 3. Recomputing the target features at each reference-enabled timestep makes both correspondence estimation and relational refinement adapt to the evolving target state.

Multiple references. Let $\mathcal{R} = \{1, \ldots, K\}$ index the reference images. For each reference $k \in \mathcal{R}$, we independently compute a refined correspondence $\boldsymbol{C}^{\star,(k)}$ and transported values

$$\boldsymbol{V}^{\text{tr},(k)} = \boldsymbol{C}^{\star,(k)} \boldsymbol{V}^{\text{ref},(k)}_{\text{img}}. \tag{13}$$

Applying equation 11 yields the corresponding full value tensor $\boldsymbol{V}^{*,(k)}$. Using the shared target attention $\boldsymbol{A}$, we combine the independently gated corrections as

$$\boldsymbol{O}^* = \boldsymbol{A}\boldsymbol{V} + \sum_{k \in \mathcal{R}} \tilde{\boldsymbol{m}}^{(k)}_{\text{place}} \odot \left(\boldsymbol{A}\boldsymbol{V}^{*,(k)} - \boldsymbol{A}\boldsymbol{V}\right). \tag{14}$$

The joint target forward produces $\boldsymbol{v}_t^{\text{tgt}|\mathcal{R}}$. At the velocity level, we use the union placement gate

$$\boldsymbol{m}^{\text{vel}}_{\cup} = \min\left(\boldsymbol{1}, \sum_{k \in \mathcal{R}} \boldsymbol{m}^{(k)}_{\text{place}}\right),$$

so that equation 3 becomes

$$\Delta \boldsymbol{v}_t = \boldsymbol{v}_t^{\text{tgt}} - \boldsymbol{v}_t^{\text{src}} + \lambda \boldsymbol{m}^{\text{vel}}_{\cup} \odot \left(\boldsymbol{v}_t^{\text{tgt}|\mathcal{R}} - \boldsymbol{v}_t^{\text{tgt}}\right). \tag{15}$$

Each additional reference requires its own reference forward, feature cache, and correspondence refinement, while the target attention sequence length remains unchanged.

Algorithm 1 summarizes the single-reference sampler. The frozen velocity predictor v_θ is the T2I model, and $\mathcal{D}$ denotes its latent decoder. We use v^*_θ to denote the same model evaluated with the cached attention intervention; RelationalRefine solves equation 9.

5 Experiments

5.1 Setup

Datasets. We evaluate RefFlow on a DreamBooth-derived benchmark (Ruiz et al., 2023) consisting of 84 cases, widely used for training-free editing (Alimohammadi et al., 2025; Song et al., 2026; Nam et al., 2024), and our newly curated RefEdit-Eval comprising 126 cases. RefEdit-Eval covers font, texture, style, object-state, and general object editing, using real Unsplash images and synthetic images generated by Qwen-Image-2512 (Qwen Team, 2025b).

Baselines. We implement RefFlow on FLUX.1-dev (Black Forest Labs, 2024) and compare against six baselines: training-free text-guided editors, FlowEdit (Kulikov et al., 2025) and FlowAlign (Kim et al., 2026); training-free reference-guided methods, SpecRef (Chen & Huang, 2023) and ReInversion (Li et al., 2025a); and trained multimodal editing models, DreamOmni2 (Xia et al., 2025) and Qwen-Image-Edit-2511 (Qwen Team, 2025a)[1].

[1] VicoEdit Song et al. (2026) is not included due to the lack of an official implementation.

Algorithm 1 REFFLOW with relational reference value transport

Input: Source/reference latents $\boldsymbol{z}^{\mathrm{src}}, \boldsymbol{z}^{\mathrm{ref}}$; editing state $\boldsymbol{y}_{t_n}$; prompt conditions $c^{\mathrm{src}}, c^{\mathrm{tgt}}, c^{\mathrm{ref}}$; masks $\boldsymbol{m}_{\mathrm{ref}}, \boldsymbol{m}_{\mathrm{place}}, \tilde{\boldsymbol{m}}_{\mathrm{place}}$; remaining schedule $\{t_k\}_{k=0}^{n}$; reference strength λ.
Output: Edited image $\boldsymbol{I}^{\mathrm{edit}}$.

for $k = n, n-1, \ldots, 1$ **do**
  $t \leftarrow t_k, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  $\boldsymbol{z}_t^{\mathrm{src}} \leftarrow (1-t)\boldsymbol{z}^{\mathrm{src}} + t\boldsymbol{\epsilon}, \qquad \boldsymbol{z}_t^{\mathrm{ref}} \leftarrow (1-t)\boldsymbol{z}^{\mathrm{ref}} + t\boldsymbol{\epsilon}$
  $\boldsymbol{z}_t^{\mathrm{tgt}} \leftarrow \boldsymbol{y}_t + \boldsymbol{z}_t^{\mathrm{src}} - \boldsymbol{z}^{\mathrm{src}}$
  Run $v_\theta(\boldsymbol{z}_t^{\mathrm{ref}}, t; c^{\mathrm{ref}})$; cache $\boldsymbol{K}_{\mathrm{img}}^{\mathrm{ref}}, \boldsymbol{V}_{\mathrm{img}}^{\mathrm{ref}}$.
  $\boldsymbol{v}_t^{\mathrm{src}} \leftarrow v_\theta(\boldsymbol{z}_t^{\mathrm{src}}, t; c^{\mathrm{src}})$
  $\boldsymbol{v}_t^{\mathrm{tgt}} \leftarrow v_\theta(\boldsymbol{z}_t^{\mathrm{tgt}}, t; c^{\mathrm{tgt}})$; cache $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$.
  for all intervention layers **do**
    Compute $\boldsymbol{C}^0$ using equation 5–equation 6.
    $\boldsymbol{C}^\star \leftarrow \boldsymbol{C}^0$
    if the layer uses relational refinement **then**
      Evaluate $\boldsymbol{G}^{\mathrm{tgt}}, \boldsymbol{G}^{\mathrm{ref}}$ and spatial weights $\boldsymbol{W}$.
      $\boldsymbol{C}^\star \leftarrow \mathrm{RelationalRefine}\left(\boldsymbol{C}^0, \boldsymbol{G}^{\mathrm{tgt}}, \boldsymbol{G}^{\mathrm{ref}}, \boldsymbol{W}\right)$
    end if
    Cache $\boldsymbol{V}^{\mathrm{tr}} \leftarrow \boldsymbol{C}^\star \boldsymbol{V}_{\mathrm{img}}^{\mathrm{ref}}$.
  end for
  $\boldsymbol{v}_t^{\mathrm{tgt|ref}} \leftarrow v_\theta^*(\boldsymbol{z}_t^{\mathrm{tgt}}, t; c^{\mathrm{tgt}})$
  $\Delta\boldsymbol{v}_t \leftarrow \boldsymbol{v}_t^{\mathrm{tgt}} - \boldsymbol{v}_t^{\mathrm{src}} + \lambda\, \boldsymbol{m}_{\mathrm{place}} \odot \left(\boldsymbol{v}_t^{\mathrm{tgt|ref}} - \boldsymbol{v}_t^{\mathrm{tgt}}\right)$
  $\boldsymbol{y}_{t_{k-1}} \leftarrow \boldsymbol{y}_t + (t_{k-1} - t)\Delta\boldsymbol{v}_t$
end for
return $\boldsymbol{I}^{\mathrm{edit}} \leftarrow \mathcal{D}(\boldsymbol{y}_0)$

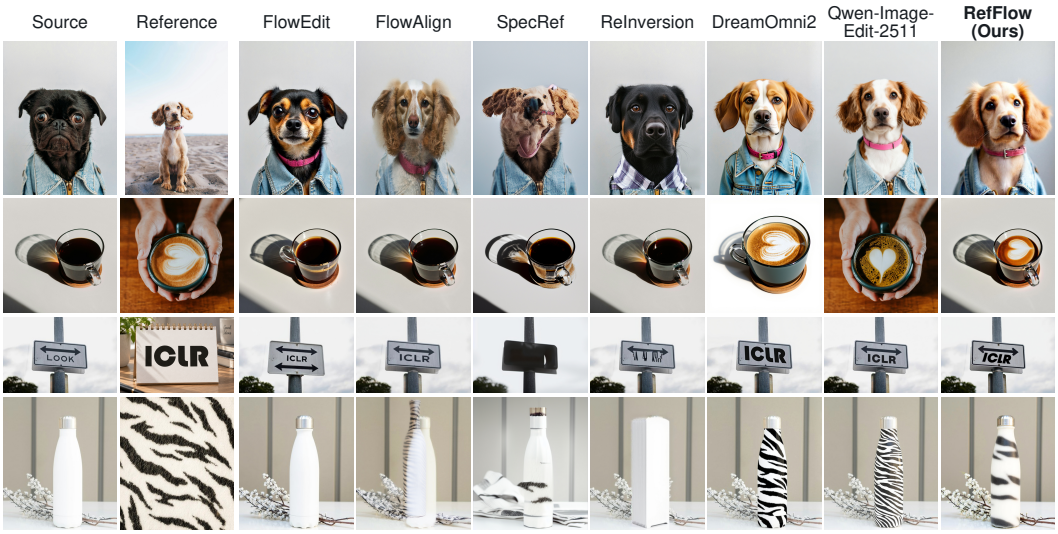

Figure 3: Qualitative comparisons on RefEdit-Eval. Rows illustrate subject adaptation, subtle attribute transfer (latte art), text and font transfer ("LOOK"→"ICLR"), and texture transfer, from top to bottom.

Metrics. Following (Kim et al., 2026; Li et al., 2025a; Song et al., 2026), we use CLIP-FG and DINO-FG for reference fidelity, CLIP-T for editing accuracy, and PSNR-BG and LPIPS-BG for source preservation. Since these metrics do not fully capture editing success (Ku et al., 2024), we additionally report Human Success Rate (Human-SR) from blind human evaluation, details are illustrated in Appendix C.

5.2 MAIN RESULTS

Results on RefEdit-Eval. Figure 3 and Table 1 show the qualitative and quantitative results:

Table 1: **Quantitative comparisons on RefEdit-Eval.** The best and second-best results are **bolded** and underlined, respectively. Cell colors indicate performance from worse to better, using per-column min–max normalization with metric directions taken into account.

		Reference fidelity		Text alignment	Background preservation		Success
Method	Category	CLIP-FG ↑	DINO-FG ↑	CLIP-T ↑	PSNR-BG ↑ (dB)	LPIPS-BG ↓	Human-SR ↑ (%)
Vanilla FlowEdit	TF-T	0.746	0.342	0.347	22.42	0.185	7.1
FlowAlign	TF-T	0.789	0.427	0.350	23.31	0.179	8.3
SpecRef	TF-R	0.809	0.366	0.336	17.20	0.319	9.5
ReInversion	TF-R	0.763	0.255	0.288	22.26	0.200	14.3
DreamOmni2	Trained-R	0.817	0.435	**0.359**	18.12	0.205	33.3
Qwen-Image-Edit-2511	Trained-R	0.839	0.478	0.349	21.14	0.193	71.4
REFFLOW (Ours)	TF-R	**0.909**	**0.686**	0.351	**25.37**	**0.175**	**90.5**

Categories: TF-T: training-free text-guided editing; TF-R: training-free reference-guided editing; Trained-R: pretrained models with learned multi-image conditioning.

Quantitatively, REFFLOW achieves the highest Human-SR and top scores in CLIP-FG, DINO-FG, PSNR-BG, and LPIPS-BG, confirming superior reference fidelity. Although CLIP-T remains slightly below the trained model, this indicates that our gains stem from faithful reference transfer rather than aggressive semantic alignment. Qualitative comparisons (Figure 3) further show that REFFLOW reliably transfers reference attributes into the correspondent source images.

Results on DreamBooth. Evaluations on the public DreamBooth dataset (Figure 4, Table 2) further validate our approach. This confirms the generalizability of residual instance transport beyond our curated benchmark. More results on various tasks are shown in Figure 9. We also evaluate the applicability of REFFLOW to another T2I backbone in Appendix F.

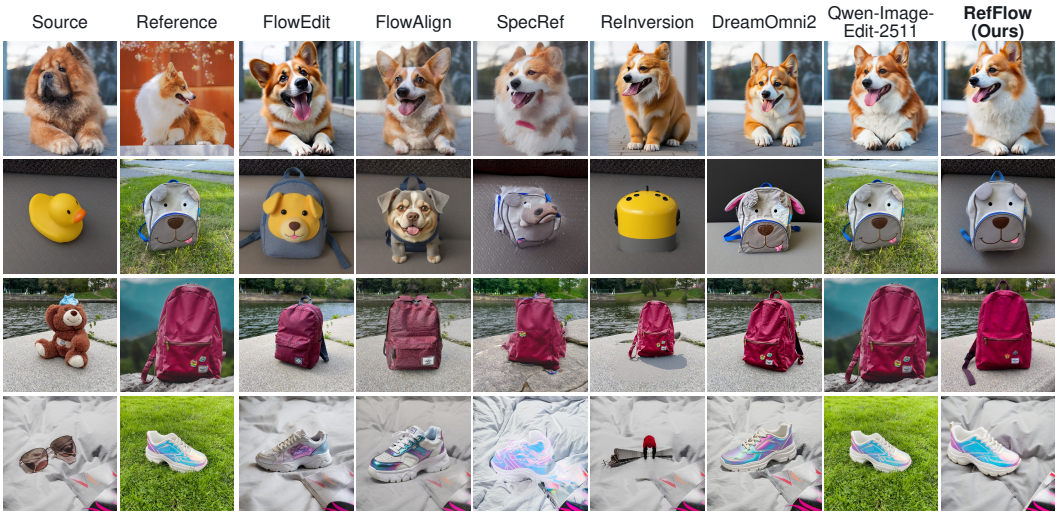

Figure 4: Qualitative comparisons on DreamBooth.

Table 2: **Quantitative comparisons on DreamBooth.** The best and second-best results are **bolded** and underlined, respectively. Colors indicate performance from worse to better, normalized independently within each metric with metric directions taken into account.

Method	Category	Reference fidelity		Text alignment	Background preservation		Success
		CLIP-FG ↑	DINO-FG ↑	CLIP-T ↑	PSNR-BG ↑ (dB)	LPIPS-BG ↓	Human-SR ↑ (%)
Vanilla FlowEdit	TF-T	0.824	0.553	0.344	18.47	0.441	5.9
FlowAlign	TF-T	0.862	0.551	0.358	19.78	0.336	8.3
SpecRef	TF-R	0.859	0.608	0.352	17.91	0.495	8.3
ReInversion	TF-R	0.790	0.490	0.288	18.70	0.380	15.5
DreamOmni2	Trained-R	0.904	0.706	**0.367**	17.91	0.328	50.0
Qwen-Image-Edit-2511	Trained-R	0.917	0.780	0.351	14.59	0.530	57.1
RefFlow (Ours)	TF-R	**0.948**	**0.896**	0.366	**20.03**	**0.324**	**66.7**

5.3 Ablation Study

We examine the strength and spatial support of the reference-induced residual in equation 3 by varying λ and enabling or disabling $\boldsymbol{m}_{\text{place}}$. Figure 5 shows that moderate reference guidance achieves the highest CLIP-FG, with both settings peaking at $\lambda = 0.5$. Further increasing the residual strength reduces reference fidelity, while spatial gating generally improves performance across the tested strengths. These results support combining moderate reference guidance with spatially restricted residual corrections. Additional ablations on spatial control and relational refinement are provided in Appendices G and H.

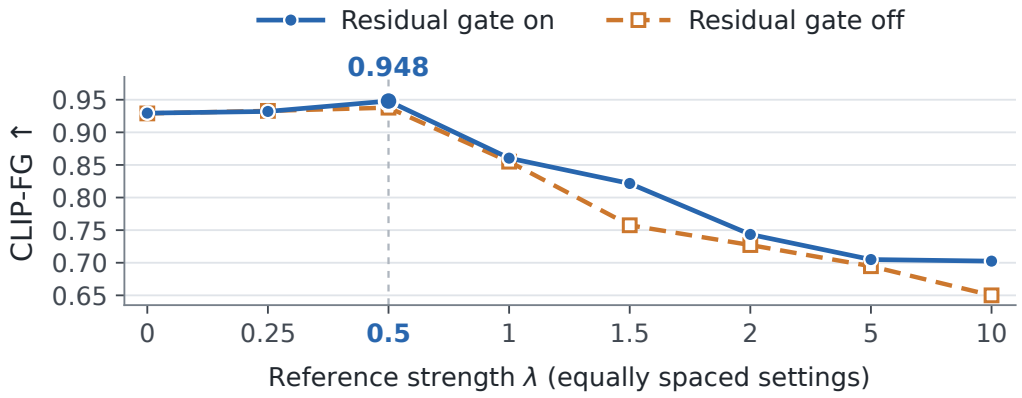

Figure 5: Effect of residual strength and spatial gating.

6 Conclusion

We presented RefFlow, a training-free framework for reference-guided image editing with a frozen T2I flow model. Our approach augments the semantic editing direction with a single reference-induced velocity residual, enabling instance-specific appearance control. Relationally Consistent Reference Value Transport adapts reference retrieval to the evolving target state and refines the resulting correspondences under within-image feature relations, while spatial gating localizes the correction to the target placement region. The framework naturally supports single- and multi-reference editing across diverse tasks. Experiments on RefEdit-Eval and DreamBooth-derived cases show that RefFlow outperforms the evaluated training-free and training-based baselines in reference fidelity and editing success, with competitive text alignment and background preservation. These results highlight residual instance transport as an effective mechanism for introducing visual reference control into pretrained generative flows without additional training.

AI USE STATEMENT

General-purpose generative AI assistants were used to support code development and language polishing. The authors reviewed all AI-assisted content and assume full responsibility for the paper.

REFERENCES

Amirhossein Alimohammadi, Aryan Mikaeili, Sauradip Nag, Negar Hassanpour, Andrea Tagliasacchi, and Ali Mahdavi-Amiri. Cora: Correspondence-aware image editing using few step diffusion. In *ACM SIGGRAPH 2025 Conference Papers*. Association for Computing Machinery, 2025. doi: 10.1145/3721238.3730650. URL https://doi.org/10.1145/3721238.3730650.

Black Forest Labs. FLUX.1 [dev]. Hugging Face model card, 2024. URL https://huggingface.co/black-forest-labs/FLUX.1-dev.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. URL https://arxiv.org/abs/2506.15742.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023. URL https://arxiv.org/abs/2211.09800.

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. 22598–22608, 2023.

Songyan Chen and Jiancheng Huang. SpecRef: A fast training-free baseline of specific reference-condition real image editing. In *2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, pp. 369–375. IEEE, 2023. URL https://ieeexplore.ieee.org/document/10424842/.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. AnyDoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2307.09481.

Taichi Endo, Guoqing Hao, and Kazuhiko Sumi. FlowSlider: Training-free continuous image editing via fidelity-steering decomposition. *arXiv preprint arXiv:2604.02088*, 2026. URL https://arxiv.org/abs/2604.02088.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Gemma Team. Gemma 4 Technical Report, 2026. URL https://arxiv.org/abs/2607.02770.

Runze He, Kai Ma, Linjiang Huang, Shaofei Huang, Jialin Gao, Xiaoming Wei, Jiao Dai, Jizhong Han, and Si Liu. FreeEdit: Mask-free reference-based image editing with multi-modal instruction. *arXiv preprint arXiv:2409.18071*, 2024. URL https://arxiv.org/abs/2409.18071.

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://arxiv.org/abs/2208.01626.

Lifan Jiang, Boxi Wu, Yuhang Pei, Tianrun Wu, Yongyuan Chen, Yan Zhao, Shiyu Yu, and Deng Cai. SNR-Edit: Structure-aware noise rectification for inversion-free flow-based editing. *arXiv preprint arXiv:2601.19180*, 2026. URL https://arxiv.org/abs/2601.19180.

Jeongsol Kim, Yeobin Hong, Jonghyun Park, and Jong Chul Ye. FlowAlign: Trajectory-regularized, inversion-free flow-based image editing. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2505.23145.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. VIEScore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.663. URL https://aclanthology.org/2024.acl-long.663/.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. FlowEdit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19721–19730, 2025. URL https://arxiv.org/abs/2412.08629.

Yuke Li, Lianli Gao, Ji Zhang, Pengpeng Zeng, Lichuan Xiang, Hongkai Wen, Heng Tao Shen, and Jingkuan Song. Reversible inversion for training-free exemplar-guided image editing. *arXiv preprint arXiv:2512.01382*, 2025a. URL https://arxiv.org/abs/2512.01382.

Zixiang Li, Yue Song, Jianing Peng, Ting Liu, Jun Huang, Xiaochao Qu, Luoqi Liu, Wei Wang, Yao Zhao, and Yunchao Wei. On exact editing of flow-based diffusion models. *arXiv preprint arXiv:2512.24015*, 2025b. URL https://arxiv.org/abs/2512.24015.

Liangsi Lu, Xuhang Chen, Minzhe Guo, Shichu Li, Jingchao Wang, and Yang Shi. ChordEdit: One-step low-energy transport for image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2602.19083.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2108.01073.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2211.09794.

Seungwoo Nam, Younghyun Kim, Dongyoon Wee, Tae-Kyun Kim, and Youngjung Uh. Dreammatcher: Appearance matching for diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8003–8012, 2024.

Qwen Team. Qwen-Image-Edit-2511: Model card. https://huggingface.co/Qwen/Qwen-Image-Edit-2511, 2025a. Version-specific model documentation; accessed September 8, 2026.

Qwen Team. Qwen-image-2512. https://huggingface.co/Qwen/Qwen-Image-2512, 2025b. Text-to-image model; accessed September 2026.

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. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ruiz_DreamBooth_Fine_Tuning_Text-to-Image_Diffusion_Models_for_Subject-Driven_Generation_CVPR_2023_paper.html.

Rui Song, Guo-Hua Wang, Qing-Guo Chen, Weihua Luo, Tongda Xu, Zhening Liu, Yan Wang, Zehong Lin, and Jun Zhang. Training-free image editing with visual context integration and concept alignment. *arXiv preprint arXiv:2604.04487*, 2026. URL https://arxiv.org/abs/2604.04487.

Anqi Tang, Wenhao Sun, and Zhaoqiang Liu. Wavelet-guided semantic signal compensation for inversion-free image editing. *arXiv preprint arXiv:2607.02421*, 2026. URL https://arxiv.org/abs/2607.02421.

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*, 2023. URL https://arxiv.org/abs/2211.12572.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Optimal transport for structured data with application on graphs. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6275–6284, 2019.

Bin Xia, Bohao Peng, Yuechen Zhang, Junjia Huang, Jiyang Liu, Jingyao Li, Haoru Tan, Sitong Wu, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. DreamOmni2: Multimodal instruction-based editing and generation. *arXiv preprint arXiv:2510.06679*, 2025. URL https://arxiv.org/abs/2510.06679.

Bin Xia, Bohao Peng, Yuechen Zhang, Junjia Huang, Jiyang Liu, Jingyao Li, Haoru Tan, Sitong Wu, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. DreamOmni2: Multimodal instruction-based editing and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Xia_DreamOmni2_Multimodal_Instruction-based_Generation_and_Editing_CVPR_2026_paper.pdf.

Chenxi Xie, Minghan Li, Shuai Li, Yuhui Wu, Qiaosi Yi, and Lei Zhang. DNAEdit: Direct noise alignment for text-guided rectified flow editing. *arXiv preprint arXiv:2506.01430*, 2025. URL https://arxiv.org/abs/2506.01430.

Tianrui Zhu, Shiyi Zhang, Jiawei Shao, and Yansong Tang. KV-Edit: Training-free image editing for precise background preservation. *arXiv preprint arXiv:2502.17363*, 2025. URL https://arxiv.org/abs/2502.17363.

A Attention-Based Mask Estimation

The reference and target placement masks can be estimated using the T2I backbone in early inference steps. For each reference, a reference phrase specifies the visual content to retrieve, and a target phrase specifies the concept to localize in the edited image. We extract their spatial support from the attention assigned by image queries to the corresponding text tokens.

Let $\mathcal{J}_b$ denote the phrase-token indices for branch $b \in \{\mathrm{ref}, \mathrm{tgt}\}$, and let $A_{t,ij}^{b,\ell,h}$ denote the native attention probability from image query i to token j at timestep t, layer ℓ, and head h. The phrase-associated response and its aggregated spatial heatmap are

$$\begin{aligned} a_t^{b,\ell}(i) &= \frac{1}{N_h} \sum_{h=1}^{N_h} \sum_{j \in \mathcal{J}_b} A_{t,ij}^{b,\ell,h}, \\ \boldsymbol{h}^b &= \frac{1}{|\Omega_b|\,|\mathcal{L}|} \sum_{t \in \Omega_b} \sum_{\ell \in \mathcal{L}} \frac{\boldsymbol{a}_t^{b,\ell}}{\|\boldsymbol{a}_t^{b,\ell}\|_1}. \end{aligned} \tag{16}$$

Here, N_h is the number of attention heads, $\mathcal{L}$ contains the observed layers, and Ω_b indexes the collected forward evaluations. The attention probabilities are normalized over all text and image keys before the phrase-associated entries are extracted. Spatial normalization ensures that each collected map contributes equally to the aggregated heatmap.

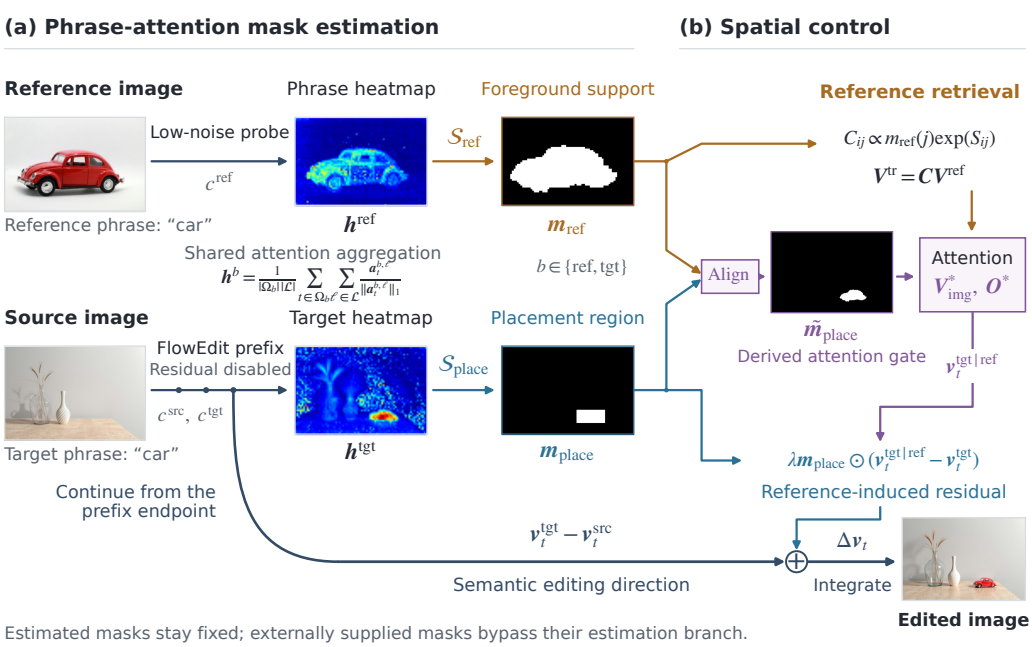

Figure 6: **Attention-based mask estimation and spatial control.** (a) A low-noise reference probe and a short reference-free FlowEdit prefix yield the reference mask $\boldsymbol{m}_{\text{ref}}$ and target placement mask $\boldsymbol{m}_{\text{place}}$ through shared phrase-attention aggregation and spatial-support extraction. (b) $\boldsymbol{m}_{\text{ref}}$ restricts reference retrieval. Aligning its foreground support with the placement region yields $\tilde{\boldsymbol{m}}_{\text{place}}$ for value intervention and attention-output gating. The original $\boldsymbol{m}_{\text{place}}$ separately gates the velocity residual added to the semantic editing direction.

For the reference branch, Ω_{ref} consists of a low-noise forward probe of the reference image conditioned on its caption. For the target branch, Ω_{tgt} contains reference-free target evaluations from a short initial FlowEdit prefix. The reference residual is disabled during this prefix, allowing the target prompt to establish the emerging editing configuration used for placement estimation.

The two masks are obtained through deterministic spatial-support extraction:

$$\boldsymbol{m}_{\text{ref}} = \mathcal{S}_{\text{ref}}(\boldsymbol{h}^{\text{ref}}), \qquad \boldsymbol{m}_{\text{place}} = \mathcal{S}_{\text{place}}(\boldsymbol{h}^{\text{tgt}}). \tag{17}$$

Both operators extract spatial support from the dominant phrase-attention response. $\mathcal{S}_{\text{ref}}$ selects the reference foreground available for retrieval, whereas $\mathcal{S}_{\text{place}}$ constructs a target placement region enclosing the localized concept. The selected reference foreground is then spatially aligned with this placement region to obtain the target-space attention gate $\tilde{\boldsymbol{m}}_{\text{place}}$. This derived gate controls value intervention and attention-output gating in equation 11 and equation 12; $\boldsymbol{m}_{\text{place}}$ controls the velocity residual in equation 3.

Figure 6 connects mask estimation to the spatial controls used by RefFlow. Both masks remain fixed after estimation; when the target placement mask is estimated, reference-guided editing resumes from the endpoint of the initial FlowEdit prefix. It is worth mentioning that this mask estimation step is not necessary when the reference and target placement masks are provided directly as external inputs.

B Implementation Details

We implement RefFlow using the frozen FLUX.1-dev (Black Forest Labs, 2024) backbone. We fix the random seed to 22 and use a 28-step sampling schedule with $n_{\max} = 23$. The guidance scales are 5.5 for the target branch and 1.5 for both the source and reference branches. The reference-induced residual strength in equation 3 is set to $\lambda = 0.5$. All source, reference, and target prompts are generated using the vision-language model Gemma 4 (Gemma Team, 2026). These settings are shared across evaluation cases. Code will be released upon acceptance of the paper.

For target-adaptive correspondence, we empirically set the matching-score scale in equation 5 to $\alpha = 4$. The relational regularization weight in equation 9 is $\beta = 2$. Value intervention is applied to all 38 single-stream transformer blocks, with relational refinement enabled only at blocks $\{9, 12, 15, 18, 21, 24, 27, 30\}$, using zero-based indexing. At each reference-enabled timestep, these blocks refine $\boldsymbol{C}^0$ across all attention heads, while the remaining intervention blocks use $\boldsymbol{C}^0$ directly.

The foreground row set is $\mathcal{F} = \{i : \tilde{m}_{\text{place}}(i) > 10^{-3}\}$. We select up to 48 landmarks at deterministically spaced indices in the flattened foreground-token ordering. The pair set $\mathcal{P}$ contains the non-self pairs between foreground rows and these landmarks, with $w_{i\ell}$ and $\boldsymbol{W}$ computed according to equation 8. Starting from $\boldsymbol{C}^0$, we perform at most three mirror-descent iterations to obtain $\boldsymbol{C}^\star$. Each iteration starts with a shared step size of 1 and allows at most five backtracking trials, halving the step size after each rejection. A candidate is accepted when the objective in equation 9, averaged over attention heads, does not increase; otherwise, the solve terminates early. When $|\mathcal{F}| < 2$, we retain $\boldsymbol{C}^0$. The refinement is performed in float32 and updates only $\boldsymbol{C}$, keeping cached features, editing latents, and model parameters fixed.

C Human Evaluation Protocol

Evaluation procedure. Three raters independently assess each edited output under blind conditions. Each item contains the source image, reference image(s), prompts, and one anonymized result. Method identities and other raters' judgments are hidden, and presentation order is randomized independently for each rater. All methods are evaluated on the same cases using the same rubric. Figure 7 illustrates the interface.

Success criteria. Raters assign *Pass* only when the result satisfies all three criteria: it fulfills the requested edit, retains task-relevant reference appearance, and preserves unrelated source content without major visual artifacts. Reference fidelity is judged relative to the intended task, allowing changes in pose, scale, and illumination needed to fit the source scene. Violation of any criterion yields *Fail*.

Aggregation and disagreements. Let $h_{ir} \in \{0, 1\}$ denote rater r's judgment for output i, where 1 indicates *Pass*. We use a predefined majority rule:

$$s_i = \mathbf{1}\left[\sum_{r=1}^{3} h_{ir} \geq 2\right], \qquad \text{Human-SR} = \frac{100}{N}\sum_{i=1}^{N} s_i. \tag{18}$$

Here, N is the number of evaluated outputs for each method. Two or three passing votes count as success; all other outcomes count as failure. Disagreements are resolved by this rule without discussion-based revision of the original votes, which are retained for analysis.

D Runtime and Memory Analysis

Figure 8 compares inference time and peak GPU memory across reference-guided editing methods. All efficiency measurements are conducted on a single NVIDIA RTX PRO 6000 GPU. To ensure a fair comparison, we include only methods that support reference-image input. Compared with SpecRef and ReInversion, RefFlow incurs additional computational cost but achieves substantially higher reference fidelity and editing success (Table 2). It also requires less inference time and GPU memory than DreamOmni2 and Qwen-Image-Edit-2511. Overall, these results demonstrate a favorable trade-off between editing quality and computational cost. The additional model evaluations required by relational correspondence refinement contribute to the inference overhead, but they enable more faithful and consistent reference transfer.

E Additional Qualitative Results

Figure 9 extends the examples in Figure 1 to additional source images, reference appearances, and editing tasks. Alongside diverse single-reference edits, we demonstrate multi-reference composition with up to five visual references using the same training-free framework.

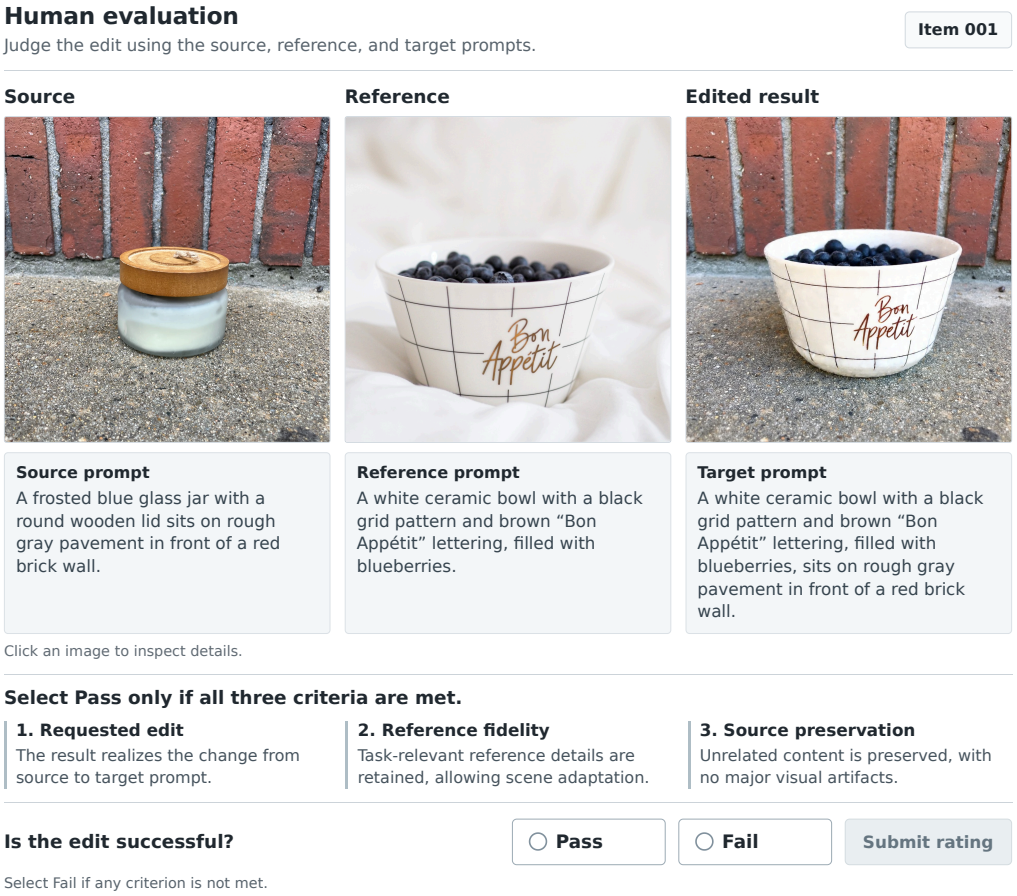

Figure 7: **Illustrative interface for blind human evaluation.** Raters inspect the prompts, source, reference, and anonymized result, then select Pass or Fail using a shared rubric.

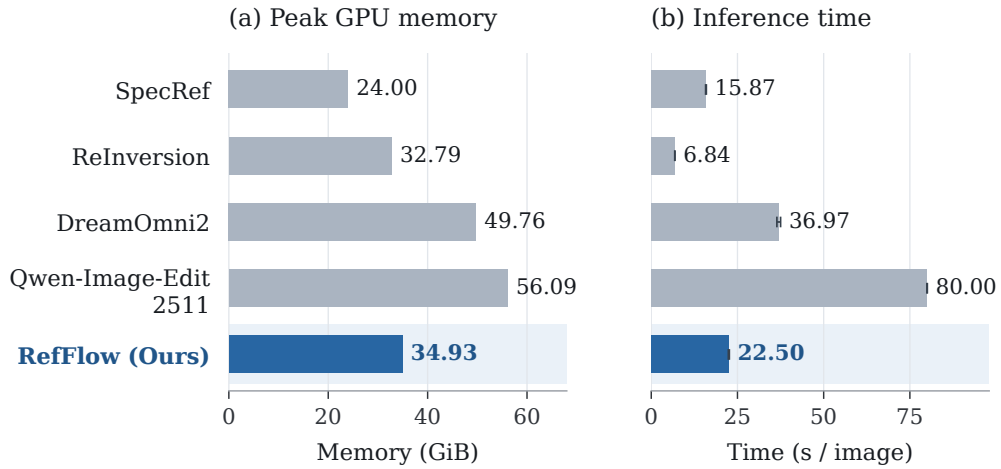

Figure 8: **Computational cost of reference-guided editing.** (a) Peak GPU memory and (b) inference time per image under the evaluated settings.

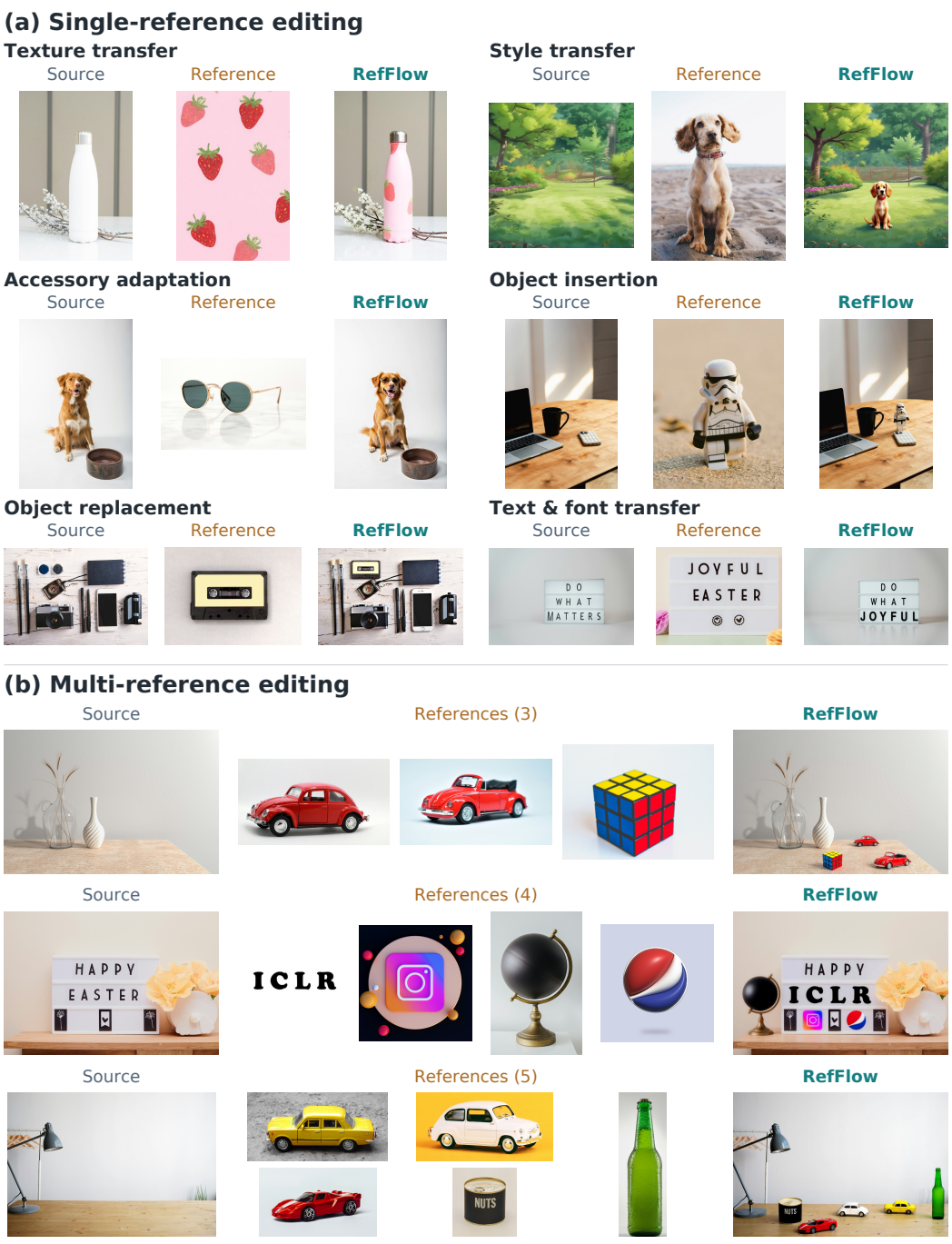

Figure 9: **Additional qualitative results of REFFLOW.** (a) Single-reference editing across various tasks. (b) Multi-reference composition with three, four, and five visual references. Each example shows the source image, reference image(s), and our edited result.

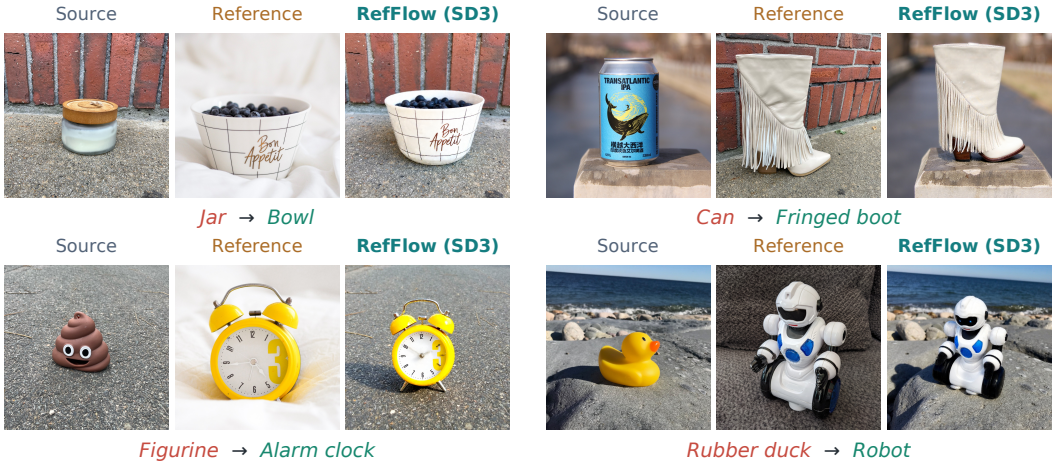

Figure 10: **Reference-guided object replacement with SD3.** Each example shows the source, reference, and REFFLOW result. Keywords below each example summarize the intended replacement.

F APPLICABILITY TO OTHER T2I BACKBONES

REFFLOW operates on predicted velocities and native attention features, allowing its formulation to be instantiated with other flow-based T2I backbones without updating model parameters. We apply REFFLOW to Stable Diffusion 3 (SD3) (Esser et al., 2024). Figure 10 shows reference-guided object replacement with SD3, transferring reference-specific appearance while largely preserving the surrounding source scene. These qualitative examples support its applicability beyond the main FLUX backbone.

G ABLATION OF SPATIAL CONTROL

We examine the roles of spatial controls and their sensitivity to mask localization errors.

G.1 EFFECTS OF SPATIAL GATING AND PLACEMENT

Ablation settings. Figure 11 examines attention-level and velocity-level spatial gating. All variants use the same experiment setting except the factor being examined. We compare two interventions:

- **No attention gate.** We set $\tilde{m}_{\text{place}}$ to one at image-token positions only in equation 11 and equation 12, retaining zero gate entries at text positions. The velocity residual remains gated by $\boldsymbol{m}_{\text{place}}$.
- **No spatial gates.** In addition to removing attention gating, we replace $\boldsymbol{m}_{\text{place}}$ with all-ones support only in equation 3. This removes both levels of gating while retaining the original placement geometry.

Findings. Removing attention gating degrades background preservation. Removing residual gating as well produces a larger increase in LPIPS-BG, while the two variants show similar changes in DINO-FG. This indicates that residual gating provides additional background protection when attention gating is absent, supporting the use of both gates to localize reference corrections.

G.2 SENSITIVITY TO MASK LOCALIZATION

We examine the effect of mask localization errors by separately translating the reference mask $\boldsymbol{m}_{\text{ref}}$ and target placement mask $\boldsymbol{m}_{\text{place}}$. Translation magnitudes are normalized by the shorter side of each mask's bounding box. As shown in Figure 12, small translations of 5% produce modest changes

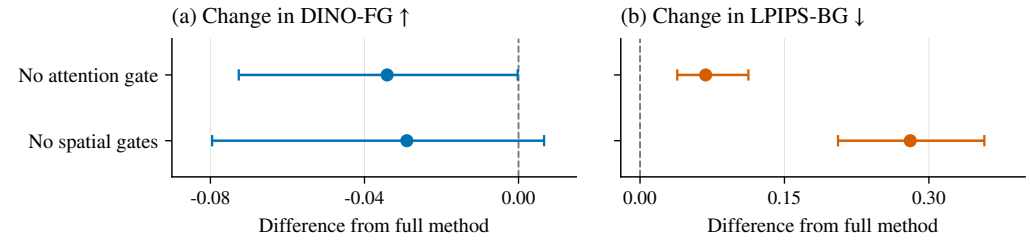

Figure 11: **Ablation of spatial gating.** Paired metric changes relative to the full method. Negative changes in DINO-FG and positive changes in LPIPS-BG indicate degradation. Error bars denote 95% confidence intervals obtained by bootstrapping cases; dashed lines indicate zero change.

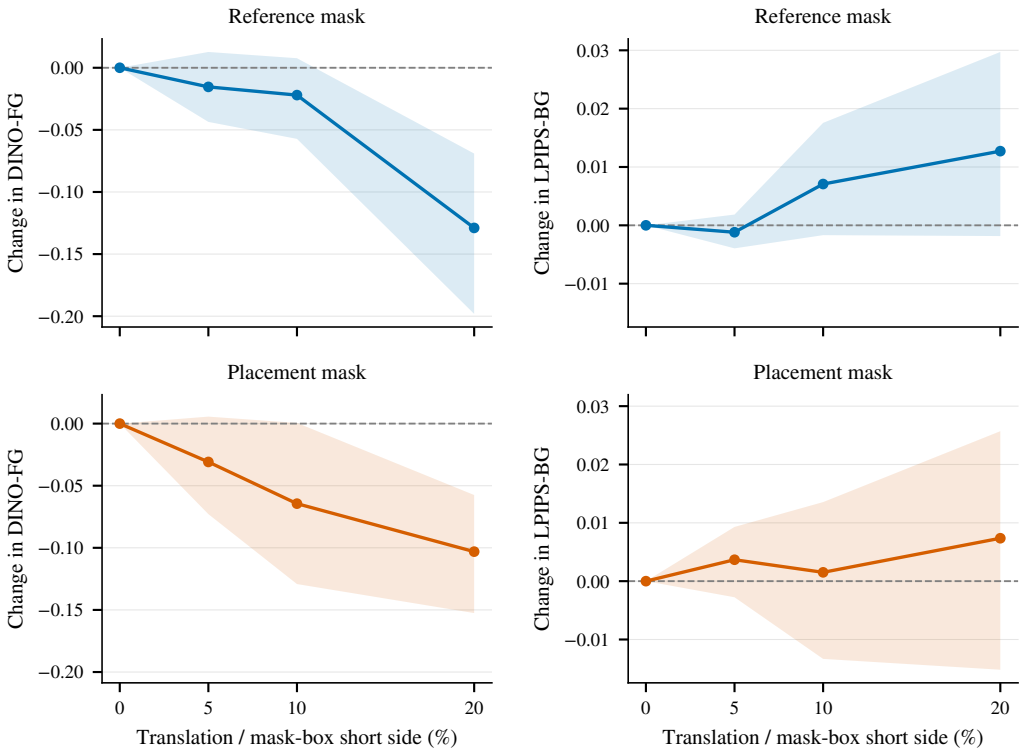

Figure 12: **Sensitivity to mask shifts.** Changes relative to the unperturbed masks when translating the reference mask (top) or placement mask (bottom). Translation is measured as a percentage of the corresponding mask bounding box's shorter side. Negative changes in DINO-FG and positive changes in LPIPS-BG indicate degradation; dashed lines denote zero change.

in reference fidelity and background preservation. Larger offsets lead to greater reductions in DINO-FG, particularly at 20%, while mean LPIPS-BG changes remain relatively small. These results suggest that the method tolerates small mask localization errors, whereas substantial misalignment compromises reference fidelity.

H Ablation of Relational Correspondence Refinement

Settings. We evaluate refinement(equation 9) on the DreamBooth subset. The *w/o refinement* variant bypasses the solver and uses $\boldsymbol{V}^{\mathrm{tr}} = \boldsymbol{C}^{0}\boldsymbol{V}^{\mathrm{ref}}$, equivalent to setting $\beta = 0$. Value intervention, spatial gating, and the reference-induced velocity residual are retained.

Table 3: **Relational correspondence refinement on DreamBooth.**

Variant	CLIP-FG ↑	DINO-FG ↑	CLIP-T ↑	PSNR-BG ↑ (dB)	LPIPS-BG ↓	Human-SR ↑ (%)
w/o refinement	0.922	0.755	0.361	19.95	0.336	50.0
w/ refinement	**0.948**	**0.896**	**0.366**	**20.03**	**0.324**	**66.7**

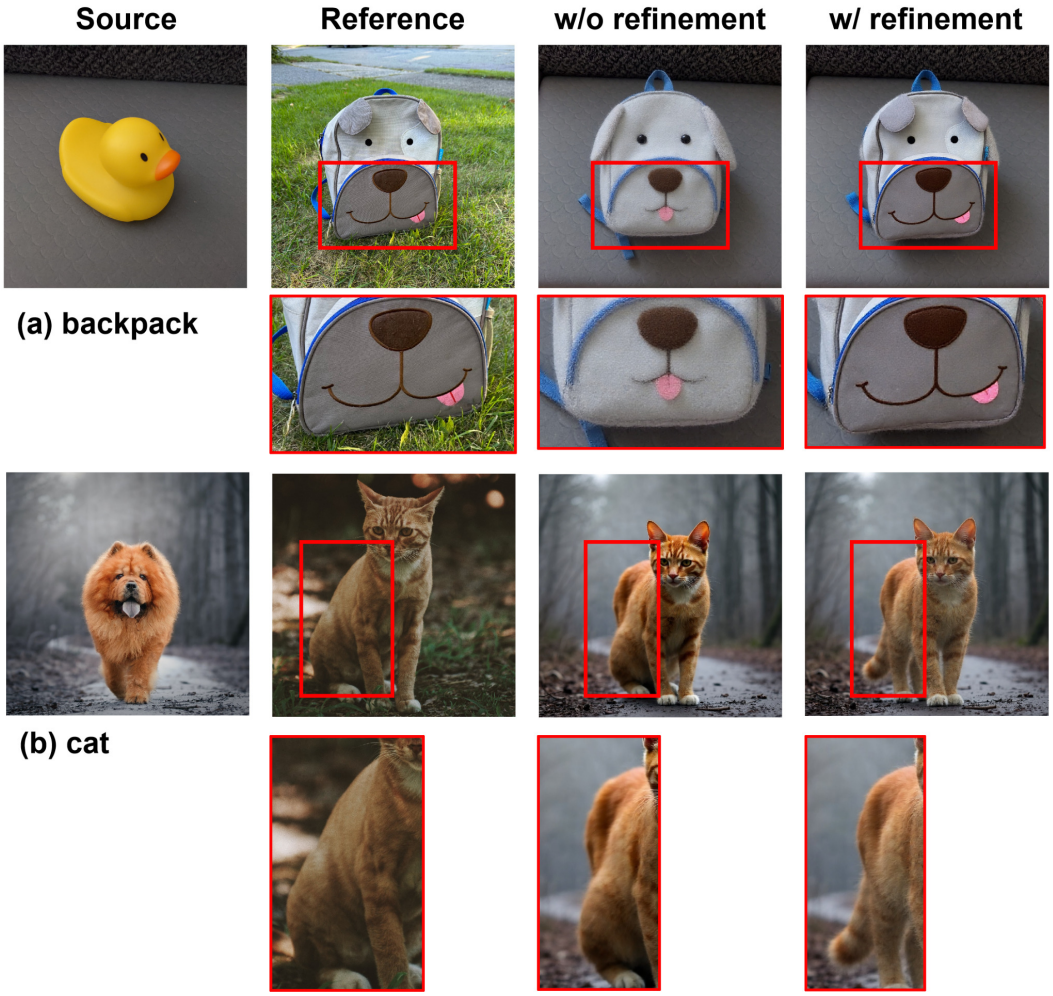

Figure 13: **Effect of relational correspondence refinement.** Two examples compare outputs with and without refinement. Boxes mark enlarged regions.

Results. Table 3 shows higher reference fidelity with refinement. Figure 13 illustrates the appearance differences, particularly the better-preserved asymmetric patches and pocket design of the backpack. These observations are consistent with more coherent reference image adaption.

I Limitations

Like other reference-guided methods (Chen & Huang, 2023; Li et al., 2025a), RefFlow relies on reference and placement masks, and masking accuracy can still affect the editing result, see Figure 12. Moreover, the refinement introduces a necessary but lightweight optimization, which can modestly reduce inference speed. For challenging tasks such as typography transfer, failures can still occur in extremely fine-grained regions, such as thin strokes, small serifs, and closely spaced glyph components, as illustrated in Figure 14. Complex edits requiring subtle contextual understanding may also

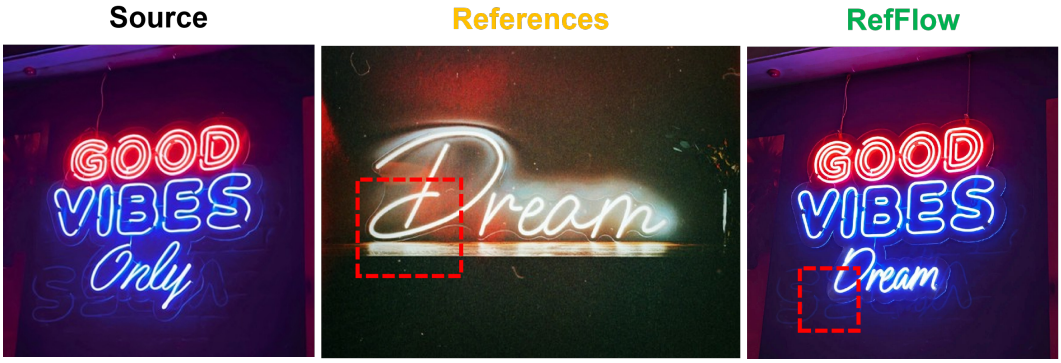

Figure 14: **Failure case in typography transfer.** The letter "D" differs slightly from the reference.

remain challenging, as our method relies on the frozen backbone's capabilities without task-specific training.