REFERENCE-IMAGE-GUIDED RELATION EDITING VIA TOKEN BOTTLENECKING AND SEMANTIC SHIFT ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing relation-centric image editing methods mainly rely on textual or spatial priors, which often fail to capture fine-grained relational configurations such as contact locations and relative poses. Compared with textual or spatial guidance, a reference image provides more fine-grained visual relation cues. Therefore, we focus on Reference-Image-Guided Relation Editing, which transfers the relation depicted in a single reference image to a source image. However, this task presents two challenges. (i) Appearance information is entangled with relation semantics, making it difficult to isolate transferable relation semantics from the reference image. This issue is particularly pronounced in unified multimodal models, where the target generation process can directly attend to reference visual representations. As a result, the generated image relies more on reference appearance than on the intended relation cues. (ii) Relation editing is prone to unintended relation-irrelevant semantic shifts that progressively accumulate during denoising because such edits require coordinated pose and spatial adjustments, whose representations in generative models are coupled with identity, background, and style. To address these challenges, we propose Relation-Aware Consistency-Preserving Editing (RACE), which (i) introduces Relation-Aware Token Bottleneck Attention (RTBA) with a limited-capacity relation-token bottleneck that induces competition between relation and appearance information to promote relation-appearance disentanglement; and (ii) proposes Cross-Modal Semantic Shift Alignment (CSSA) with a Semantic Shift Predictor (SSP). CSSA guides editing toward the intended relation shift by aligning the editing-induced image semantic shift with the intended relation-text semantic shift. Additionally, we construct a 35K-scale Relation Transfer Dataset (RTD-35K) for training. Extensive experiments demonstrate that RACE achieves superior relation transfer accuracy and source-image consistency, while showing applicability to more challenging settings, including multi-pair relation editing and cross-category relation transfer. Our code will be publicly available.

1 INTRODUCTION

Recent generative models have achieved remarkable progress in high-quality image synthesis (Rombach et al., 2022; Podell et al., 2024; Esser et al., 2024; Li et al., 2024; Zheng et al., 2024), enabling flexible image editing of object attributes and appearance (Hertz et al., 2022; Mokady et al., 2023; Kawar et al., 2023; Brooks et al., 2023; Sheynin et al., 2024; Yu et al., 2025). Beyond editing individual objects, many practical applications, such as advertising design and visual content creation, require users to modify relations among multiple objects within a scene. However, editing object relations remains relatively underexplored because relation semantics are abstract and tightly coupled with object appearance, pose, spatial configuration, and fine-grained contact patterns.

Such relation edits can be handled by general-purpose image editing methods (Brooks et al., 2023; Zhang et al., 2023b; Wu et al., 2025a;b; Labs et al., 2025; Liu et al., 2025), but these models often lack precise relation control and may alter relation-irrelevant source content. To better handle relation editing, recent works (Xu et al., 2025a; Hoe et al., 2025; 2026) introduce more structured relation modeling. Nevertheless, these methods still rely on relation guidance from coarse textual

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.

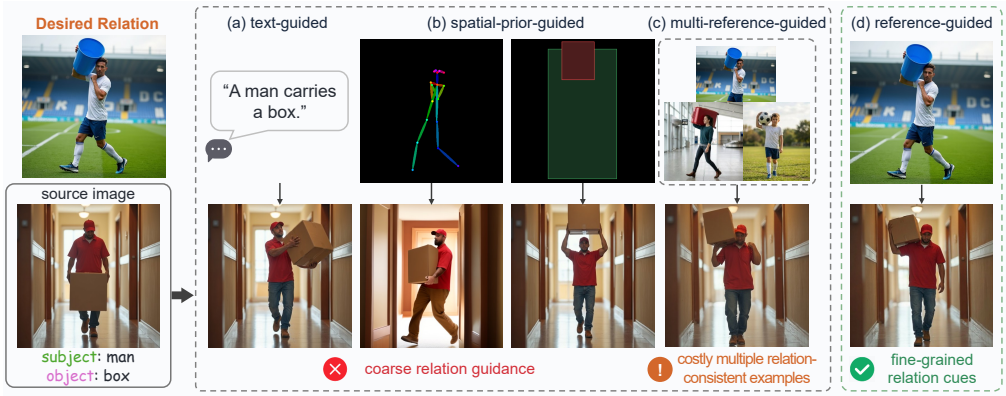

Figure 1: **Comparison of relation guidance for the "carry" example.** (a) Text specifies only a coarse relation and misses fine-grained cues such as shoulder support and hand–object contact. (b) Pose and layout priors provide only approximate spatial constraints and are sensitive to object size and shape. (c) Multi-reference guidance requires multiple relation-consistent examples that are difficult to obtain. (d) A single reference image provides fine-grained guidance for relation editing.

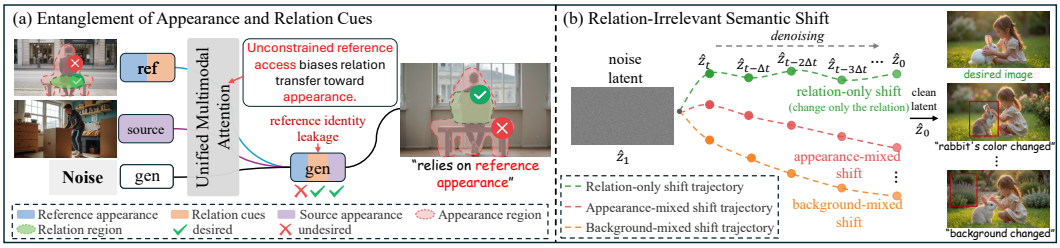

Figure 2: **Challenges of Reference-Image-Guided Relation Editing.** (a) Unified multimodal models (Deng et al., 2025; Wu et al., 2025b;a; Labs et al., 2025) allow target denoising tokens to directly attend to reference visual tokens, causing relation transfer to depend on reference appearance. (b) Relation editing methods (Xu et al., 2025a; Hoe et al., 2025; 2026) accumulate relation-irrelevant semantic shifts along the target-image denoising trajectory. Notations 'ref' and 'gen' denote reference and generated images. Red boxes highlight regions with relation-irrelevant semantic changes.

descriptions, spatial priors such as skeletons and layout boxes, or multiple relation-consistent examples, all of which have distinct limitations in capturing fine-grained visual relations. Specifically, textual guidance (Brooks et al., 2023; Wu et al., 2025a;b; Labs et al., 2025; Liu et al., 2025; Hoe et al., 2025; 2026) is often insufficient to specify fine-grained relational details. For example, as shown in Fig. 1(a), the relation "carry" may involve one or both hands, the left or right hand, and different hand–object contact locations. Spatial priors (Zhang et al., 2023c; Mou et al., 2024; Li et al., 2023; Wang et al., 2025; Shi et al., 2025; Hu et al., 2026; Hoe et al., 2026; Shakker-Labs, 2025) provide only coarse geometric cues and can be sensitive to object shape, size, and category, as shown in Fig. 1(b). Multiple-reference methods (Huang et al., 2024; Xu et al., 2025a) require costly relation-consistent examples, as shown in Fig. 1(c). Therefore, to avoid coarse textual or spatial descriptions and the need to collect multiple reference images, we focus on Reference-Image-Guided Relation Editing, which uses a single reference image to guide relation editing while preserving relation-irrelevant source content, as shown in Fig. 1(d).

This reference-image-guided formulation is well suited to unified multimodal generative models (Deng et al., 2025; Xiao et al., 2025; Wu et al., 2025b; Xie et al., 2025), which can jointly process reference and source visual representations within a unified architecture. However, effectively realizing this formulation presents two major challenges. (i) Unified multimodal models allow target denoising tokens to directly attend to reference visual tokens, and facilitate content preservation in general-purpose editing, while such unconstrained access to the full reference representation makes it difficult to isolate transferable relation cues from appearance information, since visual relation cues are highly abstract and entangled with dominant appearance features in the reference

image. As a result, the generated image relies more on reference-object appearance than on the intended relation cues, as shown in Fig. 2(a). (ii) Relation editing requires coordinated adjustments to object poses, contact states, and spatial configurations (Hoe et al., 2026), whose representations in generative models are coupled with identity, background, and style (Ma et al., 2018; Zhang et al., 2025). This coupling can therefore introduce unintended relation-irrelevant semantic shifts in identity, background, or style, which accumulate during denoising, as shown in Fig. 2(b).

To address these challenges, we propose Relation-Aware Consistency-Preserving Editing (RACE), a unified framework for Reference-Image-Guided Relation Editing. Specifically, RACE addresses these challenges with two key designs. (i) To mitigate the model's dependence on reference appearance, we introduce Relation-Aware Token Bottleneck Attention (RTBA), which uses a limited-capacity learnable relation token as a bottleneck to encourage competition between relation and appearance information, thereby promoting relation–appearance disentanglement. (ii) To suppress relation-irrelevant semantic shifts, we further propose Cross-Modal Semantic Shift Alignment (CSSA) with a Semantic Shift Predictor (SSP), which estimates editing-induced image semantic shifts from the denoising process and aligns them with the intended relation-specific textual semantic shift, thereby preserving source identity, appearance, background, and style. To support the training of RACE, we additionally construct a 35K-scale paired Relation Transfer Dataset (RTD-35K) covering 235 relation categories, with reference–target pairs sharing the same relations and source–target pairs preserving consistent appearance under different relations.

Our contributions are summarized as follows:

- We introduce Relation-Aware Token Bottleneck Attention (RTBA), which uses a limited-capacity learnable relation-token bottleneck to induce competition between relation and appearance information, thereby promoting relation–appearance disentanglement.
- We introduce Cross-Modal Semantic Shift Alignment (CSSA) with a Semantic Shift Predictor (SSP) to estimate editing-induced semantic shifts from the denoising process and suppress relation-irrelevant semantic shifts, thereby preserving source identity, appearance, background, and style.
- We construct RTD-35K, a large-scale paired dataset for Reference-Image-Guided Relation Editing. Extensive experiments validate the effectiveness of RACE across relation transfer, source-image preservation, and visual quality.

2 Related work

Multi-object Generation. Recent text-to-image models enable high-fidelity and controllable synthesis (Rombach et al., 2022; Podell et al., 2024; Esser et al., 2024; Li et al., 2024; Zheng et al., 2024), while personalized generation supports composing multiple user-specified subjects (Wang et al., 2025; Huang et al., 2025; Wu et al., 2025c). For multi-object relation synthesis, prior works impose structural constraints (Hoe et al., 2024; Shi et al., 2025; Huang et al., 2026) or extract relation representations from exemplar images (Xu et al., 2025b; Ge et al., 2024) to provide more explicit control over inter-object relations.

Relation-centric Image Editing. General-purpose image editing models can modify object relations through textual instructions (Brooks et al., 2023; Zhang et al., 2023b; Wu et al., 2025a;b; Labs et al., 2025; Liu et al., 2025), while recent methods further introduce structured relation modeling (Xu et al., 2025a; Hoe et al., 2025; 2026; Wang et al., 2024). However, textual guidance often misses fine-grained relation cues (Hoe et al., 2025; 2026), pose and layout priors provide only approximate spatial constraints (Zhang et al., 2023c; Mou et al., 2024; Li et al., 2023; Wang et al., 2025; Shi et al., 2025; Hu et al., 2026; Hoe et al., 2026; Shakker-Labs, 2025; Wang et al., 2024), and multi-reference guidance requires difficult-to-obtain relation-consistent examples (Huang et al., 2024; Xu et al., 2025a). In contrast, Reference-Image-Guided Relation Editing uses a single visual reference to provide fine-grained relation guidance for editing.

3 Method

Task definition. We focus on Reference-Image-Guided Relation Editing. Given a reference image I_{ref}, a source image I_{src}, and text descriptions specifying the subject–object pairs $(S_{\mathrm{ref}}, O_{\mathrm{ref}})$ and

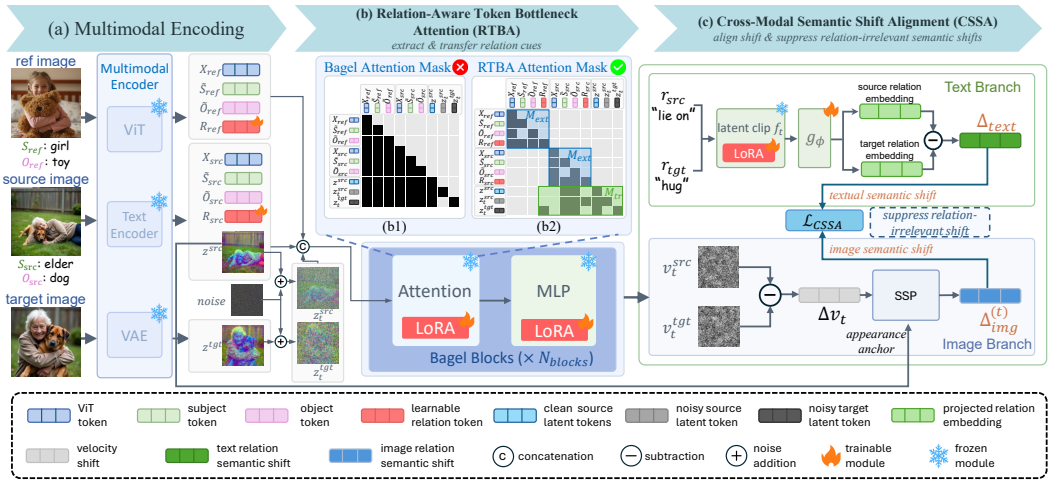

Figure 3: Overview of RACE for Reference-Image-Guided Relation Editing. Relation-Aware Token Bottleneck Attention (RTBA) extracts and routes reference relation cues through relation-token bottlenecks, reducing relation dilution and reference-image leakage. Cross-Modal Semantic Shift Alignment (CSSA) with Semantic Shift Predictor (SSP) constrains non-relational semantic shifts along the iterative generation trajectory.

$(S_{\text{src}}, O_{\text{src}})$, our goal is to generate a target image I_{tgt} that preserves the identities of the subject and object in I_{src} while modifying their relation to match that in I_{ref}.

3.1 35K-SCALE RELATION TRANSFER DATASET

Real-world image collections rarely provide sufficiently diverse image pairs that exhibit the same visual relation across different object instances while supporting relation editing. To address this, we construct a relation transfer dataset for training, as illustrated in Fig. 19. Each sample contains a reference image I_{ref}, a source image I_{src}, a target image I_{tgt}, and their corresponding textual subject–relation–object triplets $(S_{\text{ref}}, r_{\text{ref}}, O_{\text{ref}})$, $(S_{\text{src}}, r_{\text{src}}, O_{\text{src}})$, and $(S_{\text{tgt}}, r_{\text{tgt}}, O_{\text{tgt}})$.

We construct RTD-35K in three stages. (i) To obtain broadly covered relation categories and descriptions for image synthesis, we use GPT-5-mini, DeepSeek, and Qwen (OpenAI, 2025; Xu et al., 2026; Yang et al., 2025) to collect candidate relations, and GPT-5-mini (OpenAI, 2025) to describe the reference, source, and target images. The reference and target share the same relation but use different subjects and objects, while the source preserves the target appearance and background with a different relation. (ii) To construct image pairs with consistent relations between the reference and target and consistent appearance between the source and target, we synthesize the reference image with FLUX.1-dev and generate the target image using the extracted reference skeleton with a ControlNet-based pipeline. We then construct the source image by editing the target image with Qwen-Image-Edit to introduce a different relation while preserving its appearance. (iii) To ensure data quality, we use GPT-5-mini-based filtering to retain samples with consistent reference–target relations and source–target appearance. RTD-35K contains 35K reference–source–target triplets spanning 235 relation categories. More details are provided in Supplementary Sec. F.

3.2 OUR MODEL

We build RACE upon Bagel (Deng et al., 2025), a unified multimodal model with strong image understanding and generation capabilities. As shown in Fig. 3(a), following Bagel's encoding scheme (Deng et al., 2025), the reference and source images are encoded by the ViT into semantic visual tokens X_{ref} and X_{src}, respectively. The source and target images are further encoded by the VAE into latent representations z^{src} and z^{tgt}. The subject and object descriptions corresponding to the reference image are encoded into textual tokens $\widetilde{S}_{\text{ref}}$ and $\widetilde{O}_{\text{ref}}$, while those corresponding to the source image are encoded into $\widetilde{S}_{\text{src}}$ and $\widetilde{O}_{\text{src}}$.

During training, RACE performs two complementary tasks. The first is a target image generation task, where the model extracts the relation from the reference image and transfers it to the subject–object pair in the source image to generate the target image. Following flow matching (Lipman et al., 2022), z^{tgt} is interpolated with Gaussian noise to construct the noisy latent z_t^{tgt} at timestep t. In addition, with probability p_{src}, we perform a source reconstruction task, where the original relation is extracted from the source image and transferred back to the same source image for reconstruction. Similarly, z^{src} is noised to obtain z_t^{src} at timestep t.

Relation-Aware Token Bottleneck Attention. Relation cues are highly abstract and entangled with dominant appearance features, making them difficult for the model to isolate from the reference image. This issue is particularly pronounced in unified multimodal models such as Bagel, where target denoising tokens can directly access reference visual representations. Without explicit constraints, the model tends to rely on reference appearance rather than relation cues. To address this issue, we design Relation-Aware Token Bottleneck Attention (RTBA), a lightweight attention masking mechanism that uses learnable relation tokens as limited-capacity bottlenecks to disentangle transferable relation cues from reference appearance while preserving Bagel's pretrained attention structure.

We concatenate the token groups as $\mathbf{H} = \{h_1, \ldots, h_L\} = [X_{\mathrm{ref}}; \widetilde{S}_{\mathrm{ref}}; \widetilde{O}_{\mathrm{ref}}; R_{\mathrm{ref}}; X_{\mathrm{src}}; \widetilde{S}_{\mathrm{src}}; \widetilde{O}_{\mathrm{src}}; R_{\mathrm{src}}; z^{\mathrm{src}}; z_t^{\mathrm{src}}; z_t^{\mathrm{tgt}}]$, where R_{ref} and R_{src} are learnable relation tokens used to represent the relations of the designated subject–object pairs in the reference and source images, respectively. The concatenated sequence is then fed into the attention layer, where the attention is computed as

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d}} + (1 - \mathbf{B}) \cdot (-\infty)\right) V. \tag{1}$$

Here, Q, K, and V are projections of $\mathbf{H}$ into the query, key, and value spaces, respectively, and d denotes the dimension of the query and key vectors. $\mathbf{B} \in \{0, 1\}^{L \times L}$ is a binary visibility mask, where $\mathbf{B}[q, k] = 1$ allows token h_q to attend to h_k, while $\mathbf{B}[q, k] = 0$ blocks this interaction. RTBA redesigns Bagel's attention mask $\mathbf{B}$ so that reference relation information is compressed into learnable relation tokens, which serve as limited-capacity bottlenecks for target editing. The limited capacity encourages relation and appearance information to compete within the relation tokens, thereby promoting the disentanglement of transferable relation cues from reference appearance. We construct the final visibility mask as

$$\mathbf{B} = (\mathbf{M} \wedge \mathbf{C}) \vee \mathbf{S}. \tag{2}$$

Here, $\wedge$ and $\vee$ denote element-wise logical AND and OR operations, respectively. $\mathbf{C}$ is the causal mask inherited from Bagel (Deng et al., 2025), which restricts each token to attend to itself and preceding tokens. In addition, $\mathbf{S}$ is Bagel's intra-image bidirectional mask, which enables bidirectional interactions among visual tokens within the same image block. Our relation-specific structural mask $\mathbf{M}$ controls the visibility patterns required for relation extraction and transfer.

Specifically, as illustrated in Fig. 3(b2), we decompose $\mathbf{M}$ as $\mathbf{M} = \mathbf{M}_{\mathrm{ext}} \vee \mathbf{M}_{\mathrm{tr}}$, where $\mathbf{M}_{\mathrm{ext}}$ controls relation extraction and $\mathbf{M}_{\mathrm{tr}}$ controls relation transfer, corresponding to the blue and green regions, respectively. Under $\mathbf{M}_{\mathrm{ext}}$, $\widetilde{S}_n$ and $\widetilde{O}_n$ interact with X_n to localize the designated subject and object, and the learnable relation token R_n extracts the visual relation between this subject–object pair from the corresponding image, with $n \in \{\mathrm{ref}, \mathrm{src}\}$. Under $\mathbf{M}_{\mathrm{tr}}$, the clean source latent z^{src} interacts with $\widetilde{S}_{\mathrm{src}}$ and $\widetilde{O}_{\mathrm{src}}$ to perceive the subject and object whose relation should be modified. The target noisy latent z_t^{tgt} then attends to R_{ref} together with the source-side conditions z^{src}, $\widetilde{S}_{\mathrm{src}}$, and $\widetilde{O}_{\mathrm{src}}$, while direct access to the reference visual representation is blocked. The source reconstruction branch follows the same pattern using R_{src}, preserving the source relation while providing a source-side reconstruction trajectory. An intuitive illustration of the RTBA masking mechanism is provided in Supplementary Sec. B.1.

Cross-Modal Semantic Shift Alignment. Although RTBA effectively transfers reference relation cues, relation editing still requires coordinated adjustments to object poses, contact states, and spatial configurations (Hoe et al., 2026). In generative models, these structural factors are often coupled with identity, appearance, background, and style (Ma et al., 2018; Zhang et al., 2025). As a result, adjusting relation-related structures also introduces relation-irrelevant semantic shifts. Moreover, noise perturbations during iterative denoising further aggravate such unintended shifts, causing them to accumulate along the denoising trajectory, as shown in Fig. 2(b). To address this issue, we propose

Cross-Modal Semantic Shift Alignment (CSSA), which aligns the editing-induced image semantic shift with the intended textual semantic shift, as shown in Fig. 3(c). Given the source and target relation descriptions r_{src} and r_{tgt}, we encode the intended textual semantic shift as

$$\Delta_{\mathrm{text}} = g_\phi(f_t(r_{\mathrm{tgt}})) - g_\phi(f_t(r_{\mathrm{src}}))\,, \tag{3}$$

where $f_t(\cdot)$ denotes the latent-CLIP text encoder (Becker et al., 2025) and $g_\phi(\cdot)$ denotes the learnable projection network parameterized by ϕ. To estimate the image semantic shift induced by editing, a straightforward approach is to compute the semantic difference between the final generated image and the source image. However, such final-image supervision is post-hoc and cannot suppress relation-irrelevant shifts arising during the denoising process. We therefore introduce a Semantic Shift Predictor (SSP), which uses the source latent z^{src} and the trajectory-level velocity difference Δv_t between the target-editing and source-reconstruction branches to estimate the semantic shift directly from the denoising process as

$$\Delta^{(t)}_{\mathrm{img}} = SSP(z^{\mathrm{src}}, \Delta v_t, t)\,. \tag{4}$$

Here, the trajectory-level velocity difference Δv_t captures the local denoising-trajectory deviation induced by relation editing. Meanwhile, z^{src} serves as a source-content anchor that provides the identity, appearance, layout, and background context needed to interpret this latent-space deviation.

We formulate CSSA as a contrastive learning objective to align the image and textual semantic shifts. Specifically, for each sample, the predicted image semantic shift is paired with the corresponding textual semantic shift, while the textual semantic shifts of other samples in the same batch serve as negative samples. The objective is formulated as

$$\mathcal{L}_{\mathrm{CSSA}} = \mathbb{E}_{t\sim\mathcal{U}(0,1)}\left[-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\cos(\Delta^{(t)}_{\mathrm{img},i}, \Delta_{\mathrm{text},i})/\tau)}{\sum_{j=1}^{N}\exp(\cos(\Delta^{(t)}_{\mathrm{img},i}, \Delta_{\mathrm{text},j})/\tau)}\right], \tag{5}$$

where N is the batch size, i and j index samples in the batch, and τ is the temperature. During training, we also keep the source reconstruction task with probability p_{src}. This provides a stable source-side trajectory under the original relation R_{src}. The derivation of this trajectory-based estimation and additional implementation details are provided in the Supplementary Material B.2.

3.2.1 Training and Inference Strategy

Progressive training. We adopt a two-stage training strategy. In the first stage, we optimize the Bagel editing backbone without activating CSSA, allowing the target-editing and source-reconstruction denoising trajectories to become stable for semantic-shift estimation. In the second stage, we activate CSSA and jointly fine-tune the Bagel backbone, latent-CLIP text encoder f_t, relation projection network g_ϕ, and SSP. This progressive strategy avoids noisy semantic supervision from unstable semantic-shift estimates at early training stages. Detailed training objectives and algorithms are provided in Supplementary Sec. B.4.

Relation-aware Classifier-Free Guidance (R-CFG). To improve generation stability and visual quality during inference, we adopt Relation-aware Classifier-Free Guidance (R-CFG) based on relation-token conditioning (Ho & Salimans, 2022). During training, with probability p_{cfg}, we drop the relation-token condition to construct the unconditional branch. Specifically, the relation token is zeroed before each Bagel transformer block, ensuring that relation conditioning remains consistently suppressed throughout the denoising network. At inference, let $\hat{z}^{\mathrm{tgt}}_t$ denote the target latent state at timestep t along the denoising trajectory. We compute the guided velocity as

$$v^{\mathrm{tgt}}_t = v_\theta\left(\hat{z}^{\mathrm{tgt}}_t, t, c^{\mathrm{tgt}}_{\mathrm{base}}\right) + w\left(v_\theta\left(\hat{z}^{\mathrm{tgt}}_t, t, c^{\mathrm{tgt}}\right) - v_\theta\left(\hat{z}^{\mathrm{tgt}}_t, t, c^{\mathrm{tgt}}_{\mathrm{base}}\right)\right), \tag{6}$$

where $c^{\mathrm{tgt}}_{\mathrm{base}}$ denotes the target-editing condition with the relation token zeroed. The guidance scale w controls the strength of relation conditioning. This guidance strengthens relation control and improves the stability and quality of the edited results. Detailed training and inference algorithms are provided in Algorithms 1 and 2 in the Supplementary Material.

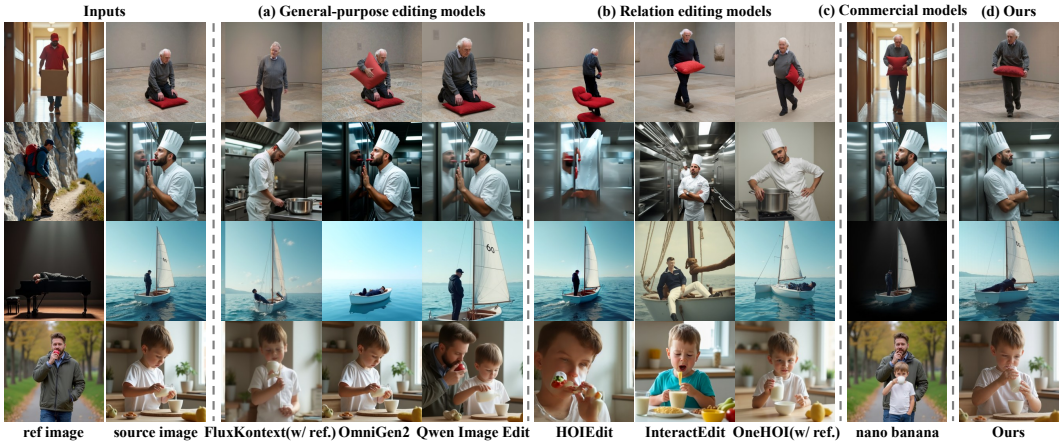

Figure 4: Qualitative comparison on Reference-Image-Guided Relation Editing (Labs et al., 2025; Wu et al., 2025b;a; Xu et al., 2025a; Hoe et al., 2025; 2026; Google, 2025).

Table 1: Quantitative comparison on HICORTBench. The best and second-best results are highlighted in bold and underlined, respectively.

Method	Objective Metrics					User Study		
	relation consistency↑	EI↑	HPSv2↑	Image Reward↑	PickScore↑	Relation Consistency↑	Editing Consistency↑	Image Quality↑
MSDiffusion (ICLR'25) (Wang et al., 2025)	0.3522	0.3578	0.2467	-0.5539	20.8465	44.00	47.27	44.36
DreamRelation (CVPR'25) (Shi et al., 2025)	0.3444	<u>0.3972</u>	0.2608	-0.0250	21.2061	51.64	48.64	50.55
DyRef (ECCV'26) (Huang et al., 2026)	<u>0.3984</u>	0.1762	0.2258	-1.1749	20.3358	53.18	29.24	29.09
FluxKontext (arXiv'25) (Labs et al., 2025)	0.2058	0.3016	0.2569	-0.4835	20.7687	42.55	<u>65.09</u>	<u>67.64</u>
FluxKontext (w/ ref.) (Labs et al., 2025)	0.3519	0.3589	0.2296	-0.7482	20.7715	41.82	44.00	31.64
OmniGen2 (arXiv'25) (Wu et al., 2025b)	0.2216	0.3229	0.2561	-0.4411	20.7239	41.36	50.91	39.09
qwen-image-edit (arXiv'25) (Wu et al., 2025a)	0.2157	0.2628	0.2518	-0.6670	20.6212	29.82	53.82	56.36
HOIEdit (TVC'25) (Xu et al., 2025a)	0.2703	0.2718	0.2083	-1.2879	20.2326	34.09	59.09	36.36
InteractEdit (arXiv'25) (Hoe et al., 2025)	0.3595	0.2824	0.2563	0.0200	21.3331	49.09	36.36	45.09
OneHOI (CVPR'26) (Hoe et al., 2026)	0.2602	0.2815	<u>0.2749</u>	<u>0.4000</u>	<u>21.3832</u>	54.09	40.45	52.27
OneHOI (w/ ref.) (Hoe et al., 2026)	0.3698	0.3637	0.2670	0.3138	21.2510	48.64	60.30	53.64
nano banana (commercial) (Google, 2025)	0.3125	0.3565	0.2422	-0.5825	20.8806	<u>73.64</u>	39.32	62.05
Ours	**0.4713**	**0.4598**	**0.2751**	**0.4564**	**21.4602**	**81.45**	**76.36**	**76.36**

4 EXPERIMENT

4.1 EXPERIMENTAL SETTINGS

Implementation Details. We use a batch size of 12, with $p_{\text{src}} = 0.2$ for source reconstruction and $p_{\text{cfg}} = 0.2$ for classifier-free guidance dropping. Training uses 6,650 warm-up steps without CSSA and 1,200 joint fine-tuning steps with CSSA. All experiments are conducted on 2 NVIDIA A100 GPUs. Additional training configurations and hyperparameter experiments are provided in Supplementary Secs. D.1 and D.2, respectively.

Benchmarks. We evaluate our method on two benchmarks. To assess generalization to real-world images, we first construct HICORTBench from HICO-DET (Chao et al., 2018), consisting of 30 relation categories and 1,000 test samples. To evaluate Reference-Image-Guided Relation Editing across diverse relation categories, we further construct RTBench by using GPT-5-mini (OpenAI, 2025) to generate reference and source image descriptions and FLUX.1-dev to synthesize the corresponding images. RTBench contains 89 relation categories and 1,500 test samples.

Evaluation metrics. We evaluate relation transfer, source-image preservation, and generation quality using objective metrics and a user study. For relation transfer, we use the final-layer PViC (Zhang et al., 2023a) outputs as interaction embeddings, and measure relation consistency between the reference and generated images by their cosine similarity. Inspired by OneHOI (Hoe et al., 2026), we report the editability-identity score (EI) to jointly evaluate relation editability and identity preservation. For generation quality, we report HPSv2, ImageReward, and PickScore (Wu et al., 2023; Xu et al., 2023; Kirstain et al., 2023). We further conduct a user study

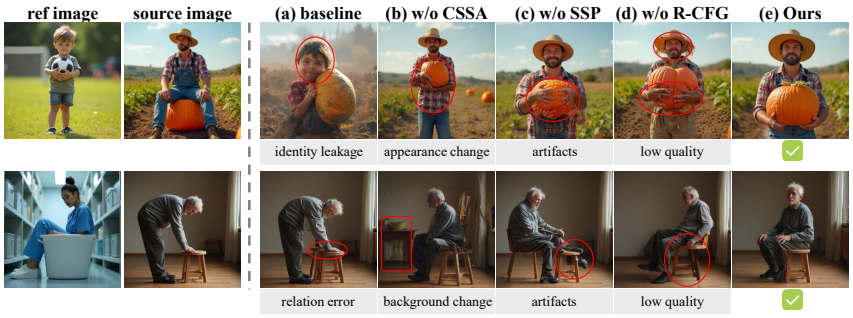

Figure 5: Qualitative ablation study of RACE. Red annotations mark failure regions.

to assess human preference for different editing results. Details of EI and the user study are provided in the Supplementary Material B.5.

4.2 Comparison with other methods

Qualitative Results. We qualitatively compare our method with state-of-the-art (SOTA) methods. For methods without reference-image input, we provide detailed textual or spatial guidance of the reference relation. For fair comparison, we adapt OneHOI and FluxKontext to accept reference images and train them under the same experimental settings as ours, denoted as OneHOI (w/ ref.) and FluxKontext (w/ ref.), respectively. As shown in Fig. 4, general-purpose methods often fail to establish the target relation or introduce artifacts. In the first-row example, FluxKontext and Qwen Image Edit fail to correctly establish the "carry" relation, while OmniGen2 introduces duplicated objects. Relation editing methods may poorly preserve source appearance, as OneHOI alters the appearance of the elderly person in the same example. Commercial models may suffer from reference-appearance leakage, with Nano Banana leaking the reference background into the generated image. In contrast, our method accurately transfers the reference relation while better preserving source content. More experimental results are provided in Supplementary Material D.7.

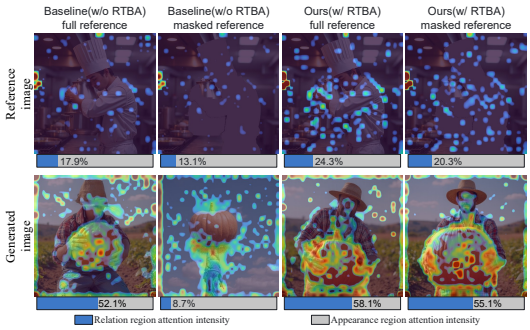

Figure 6: Attention analysis of RTBA under full and masked references. The top and bottom rows show cross-image attention on the reference and generated images, respectively. Blue and gray bars indicate relation- and appearance-region attention intensity, with higher values denoting stronger attention.

Quantitative Results. We conduct quantitative evaluation on HICORTBench. As shown in Table 1, our method improves relation consistency and editability-identity (EI) by 18.30% and 15.76% over the strongest baselines, respectively. Our method also achieves superior generation quality across HPSv2, ImageReward, and PickScore. In the user study, our method improves relation consistency, editing consistency, and image quality by 10.60%, 17.31%, and 12.89%, respectively.

4.3 Ablation studies

RTBA. We ablate RTBA, CSSA, SSP, and CFG, with qualitative and quantitative results shown in Fig. 5 and Table 2. As shown in Fig. 5(a)(b), RTBA improves relation transfer while reducing reference-specific leakage. To further analyze RTBA, we mask the object regions in the reference image and examine the cross-image attention from the generated image to the reference image. As shown in Fig. 6, after masking the reference-object

Table 2: Ablation study of our method on RTBench.

Method	relation consistency↑	EI↑	Image HPSv2↑	Image Reward↑	Image PickScore↑
Baseline	0.2770	0.3308	0.2535	0.4918	21.2388
w/o CSSA	**0.3141**	0.3708	0.2748	0.9012	21.9331
w/o SSP	0.2771	0.3417	0.2710	0.7699	21.7896
w/o R-CFG	0.1782	0.2630	0.2720	0.6773	21.7999
Ours	0.2985	**0.3851**	**0.2913**	**1.2699**	**22.4189**

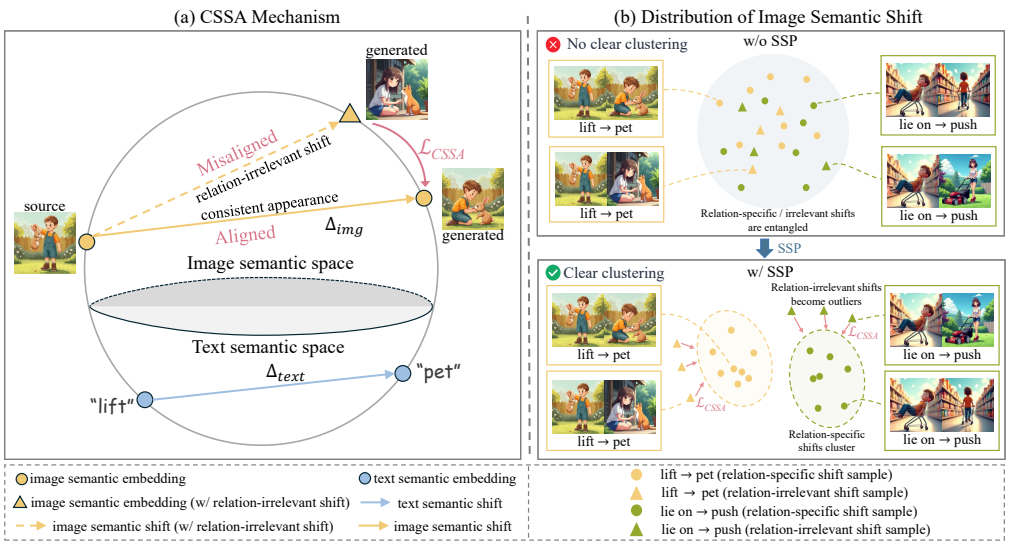

Figure 7: Mechanism and empirical analysis of Cross-Modal Semantic Shift Alignment (CSSA). (a) CSSA aligns the editing-induced image semantic shift with the intended textual relation semantic shift, guiding misaligned shifts toward the intended relation direction. (b) In the 2D PCA projection of semantic-shift embeddings, without SSP, relation-specific and relation-irrelevant semantic shifts remain entangled. With SSP, relation-specific shifts form compact clusters, while relation-irrelevant shifts appear as outliers.

regions, the generated relation regions maintain high attention intensity toward the corresponding relation regions in the reference image with RTBA, whereas this attention drops substantially without RTBA. This indicates that RTBA reduces the model's dependence on reference-object appearance.

CSSA and SSP. As shown in Fig. 5(b)(e) and Table 2, CSSA effectively suppresses relation-irrelevant semantic shifts, improving the preservation of source identity, appearance, and background. As shown in Fig. 5(c)(e) and Table 2, SSP further reduces artifacts and improves editing stability by providing more reliable trajectory-level semantic shift estimation for CSSA. To further analyze CSSA and SSP, Fig. 7(a) illustrates the cross-modal alignment mechanism of CSSA, where the editing-induced image semantic shift is aligned with the intended textual relation semantic shift to guide editing toward the intended relation direction. Fig. 7(b) further validates this mechanism by visualizing image-pair semantic-shift embeddings projected into 2D using PCA. Without SSP, relation-specific and relation-irrelevant semantic shifts remain entangled. With SSP, relation-specific shifts form compact and well-separated clusters, while relation-irrelevant shifts appear as outliers and are regularized by $\mathcal{L}_{\text{CSSA}}$.

R-CFG. Fig. 5(d)(e) and Table 2 show that removing R-CFG reduces relation consistency and generation quality, producing blurrier details and less stable edits.

5 Conclusion & Limitations

Conclusion. Relation editing has long been constrained by coarse textual descriptions and spatial priors that are sensitive to object shape, size, and category. We address this limitation by using visual references to convey relation semantics and propose RACE, where RTBA disentangles relation semantics from reference appearance and CSSA with SSP suppresses relation-irrelevant semantic shifts. We further construct RTD-35K, covering 235 relation categories. Beyond single-pair relation editing, we further extend RACE to complex scenarios such as multi-pair relation editing and cross-category relation transfer, providing more flexible relation control for applications including advertising design and visual content creation.

Limitations. RACE may still produce local artifacts in fine-grained interaction regions, such as hand–object contacts, and struggle with relations requiring substantial structural changes. Further analysis is provided in Supplementary Material E.

AI USE STATEMENT

Generative AI was used to assist in outlining, drafting, language refinement, dataset construction, and filtering for this work. All final arguments, analysis, factual verification, experimental interpretation, and editorial decisions remain solely the responsibility of the authors.

REFERENCES

Jason Becker, Chris Wendler, Peter Baylies, Robert West, and Christian Wressnegger. Controlling latent diffusion using latent clip. *arXiv preprint arXiv:2503.08455*, 2025.

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.

Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pp. 381–389. IEEE, 2018.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Mengmeng Ge, Xu Jia, Takashi Isobe, Xiaomin Li, Qinghe Wang, Jing Mu, Dong Zhou, Li Wang, Huchuan Lu, Lu Tian, Ashish Sirasao, and Emad Barsoum. Customizing text-to-image generation with inverted interaction. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 10901–10909, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3680621. URL https://doi.org/10.1145/3664647.3680621.

Google. Introducing gemini 2.5 flash image, our state-of-the-art image model. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/, 2025.

Google. Nano Banana 2: Combining pro capabilities with lightning-fast speed. The Keyword, February 2026. URL https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/. Gemini 3.1 Flash Image; published February 26, 2026; accessed August 24, 2026.

Huiguo He, Huan Yang, Zixi Tuo, Yuan Zhou, Qiuyue Wang, Yuhang Zhang, Zeyu Liu, Wenhao Huang, Hongyang Chao, and Jian Yin. Dreamstory: Open-domain story visualization by llm-guided multi-subject consistent diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jiun Tian Hoe, Xudong Jiang, Chee Seng Chan, Yap-Peng Tan, and Weipeng Hu. Interactdiffusion: Interaction control in text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6180–6189, 2024.

Jiun Tian Hoe, Weipeng Hu, Wei Zhou, Chao Xie, Ziwei Wang, Chee Seng Chan, Xudong Jiang, and Yap-Peng Tan. Interactedit: Zero-shot editing of human-object interactions in images. *arXiv preprint arXiv:2503.09130*, 2025.

Jiun Tian Hoe, Weipeng Hu, Xudong Jiang, Yap-Peng Tan, and Chee Seng Chan. Onehoi: Unifying human-object interaction generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7664–7673, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Junjie Hu, Tianyang Han, Kai Ma, Jialin Gao, Yang Song, Xianhua He, Junfeng Luo, Xiaoming Wei, and Wenqiang Zhang. Positionic: Unified position and identity consistency for image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9139–9148, 2026.

Qihan Huang, Siming Fu, Jinlong Liu, Hao Jiang, Yipeng Yu, and Jie Song. Resolving multi-condition confusion for finetuning-free personalized image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3707–3714, 2025.

Wenwang Huang, Yusen Fu, Junjie Wang, Mengfei Huang, Yulin Li, Gan Liu, Jing Cai, Yancheng He, and Zhuotao Tian. Scaling multi-reference image generation with dynamic reward optimization. *arXiv preprint arXiv:2606.26947*, 2026.

Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. Reversion: Diffusion-based relation inversion from images. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Zeyu Jiang, Lai Man Po, Xuyuan Xu, Yexin Wang, Guoping Gong, Haoxuan Wu, Chenbo Yan, Kun Li, and Yuyang Liu. Orionedit: Bridging reference and source images for generalized cross-image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9127–9138, 2026.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6007–6017. IEEE, 2023.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22511–22521. IEEE, 2023.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 99–108. IEEE, 2018.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6038–6047. IEEE, 2023.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

OpenAI. Openai gpt-5 system card. Technical Report arXiv:2601.03267, OpenAI, 2025. URL https://cdn.openai.com/gpt-5-system-card.pdf.

OpenAI. GPT Image 2 model. OpenAI API Documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-image-2. API model identifier: gpt-image-2; snapshot: gpt-image-2-2026-04-21; accessed August 24, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. URL https://api.semanticscholar.org/CorpusID:258170077.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Reve Team. Reve – ai image editor. https://app.reve.com/, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Shakker-Labs. Controlnet-union. https://huggingface.co/Shakker-Labs/FLUX.1-dev-ControlNet-Union-Pro-2.0, 2025.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *2024 ieee/cvf conference on computer vision and pattern recognition (cvpr)*, pp. 8871–8879. IEEE, 2024.

Qingyu Shi, Lu Qi, Jianzong Wu, Jinbin Bai, Jingbo Wang, Yunhai Tong, and Xiangtai Li. Dreamrelation: Bridging customization and relation generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15723–15732, 2025.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. In *International Conference on Learning Representations*, volume 2025, pp. 95118–95146, 2025.

Zhen Wang, Yilei Jiang, Dong Zheng, Jun Xiao, and Long Chen. Event-customized image generation. *arXiv preprint arXiv:2410.02483*, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18682–18692, 2025c.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13294–13304, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Tang Xu, Wenbin Wang, and Alin Zhong. Hoiedit: Human–object interaction editing with text-to-image diffusion model: T. xu et al. *The Visual Computer*, 41(8):6187–6199, 2025a.

Zhu Xu, Zhaowen Wang, Yuxin Peng, and Yang Liu. Interact-custom: Customized human object interaction image generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9500–9508, 2025b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26125–26135, 2025. doi: 10.1109/CVPR52734.2025.02433.

Frederic Z Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring predicate visual context in detecting of human-object interactions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10411–10421, 2023a.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023b.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023c.

Wei Zhang, Chenglin Zhou, Xuekang Peng, and Zhichao Lian. Inpaintingpose: Enhancing human pose transfer by image inpainting. *Image and Vision Computing*, pp. 105690, 2025.

Xiaoyan Zhang, Zechen Bai, Haofan Wang, and Yiren Song. Sigma: Selective-interleaved generation with multi-attribute tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38165–38175, 2026.

Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihan Wang, Jidong Chen, Xiaotao Gu, Yuxiao Dong, Ming Ding, and Jie Tang. Cogview3: Finer and faster text-to-image generation via relay diffusion. In *European Conference on Computer Vision*, pp. 1–22. Springer, 2024.

Table 3: Summary of the main notation used in the paper.

Symbol	Description
$I_{\mathrm{ref}}, I_{\mathrm{src}}, I_{\mathrm{tgt}}$	Reference image, source image, and target image, respectively.
S_n, O_n, r_n	Subject, object, and relation descriptions for image $n \in \{\mathrm{ref}, \mathrm{src}, \mathrm{tgt}\}$.
$X_{\mathrm{ref}}, X_{\mathrm{src}}$	Semantic visual tokens encoded from the reference and source images.
$\widetilde{S}_n, \widetilde{O}_n$	Subject and object textual tokens used for visual grounding.
$R_{\mathrm{ref}}, R_{\mathrm{src}}$	Learnable tokens representing the reference and source relations.
$\mathbf{H}, h_i, L$	Concatenated token sequence, its i-th token, and sequence length.
Q, K, V	Query, key, and value projections in attention.
d	Query/key dimension in attention.
$\mathbf{B}, \mathbf{C}, \mathbf{S}$	Final visibility mask, causal mask, and intra-image bidirectional mask.
$\mathbf{M}, \mathbf{M}_{\mathrm{ext}}, \mathbf{M}_{\mathrm{tr}}$	Relation-specific structural mask and its extraction and transfer components.
$z^{\mathrm{src}}, z^{\mathrm{tgt}}$	Clean VAE latents of the source and target images; z^{src} also provides the static source condition.
t, ϵ	Flow-matching timestep and sampled standard Gaussian noise.
$z_t^{\mathrm{src}}, z_t^{\mathrm{tgt}}$	Noisy training latents for source reconstruction and target editing.
$\hat{z}_t^{\mathrm{src}}, \hat{z}_t^{\mathrm{tgt}}$	Model-generated latent states along the two reverse trajectories.
$u_t, u_t^{\mathrm{src}}, u_t^{\mathrm{tgt}}$	Analytic flow-matching target velocities, with $u_t = \epsilon - z$ for clean latent z.
v_θ	Learned velocity predictor with parameters θ.
$v_t^{\mathrm{src}}, v_t^{\mathrm{tgt}}$	Predicted velocities along the source and target trajectories.
$c^{\mathrm{src}}, c^{\mathrm{tgt}}, c_{\mathrm{base}}^{\mathrm{tgt}}$	Source and target conditioning contexts, and target context with the relation token zeroed.
$\Delta v_t, \Delta z_0$	Target-minus-source differences in trajectory velocity and generated endpoint latent, respectively.
f_t, g_ϕ	Latent-CLIP text encoder and learnable projection network with parameters ϕ.
Δ_{text}	Intended textual semantic shift from the source to the target relation.
SSP	Semantic Shift Predictor estimating the image semantic shift from z^{src}, Δv_t, and t.
$\Delta_{\mathrm{img}}^{(t)}, \Delta_{\mathrm{img}}^{*}$	Predicted image semantic shift at timestep t and ideal endpoint semantic shift.
δ_{irr}	Relation-irrelevant semantic shift in identity, appearance, background, or style in the timestep-wise alignment example.
$f_{\mathrm{img}}, J_{f_{\mathrm{img}}}$	Image-side semantic encoder and its Jacobian.
$\mathcal{L}_{\mathrm{FM}}, \mathcal{L}_{\mathrm{src}}, \mathcal{L}_{\mathrm{tgt}}$	Flow-matching objective and its source-reconstruction and target-editing branch losses.
$\mathcal{L}_{\mathrm{CSSA}}, \lambda_{\mathrm{CSSA}}$	Contrastive semantic-shift alignment loss and its weight.
$\mathcal{L}_{\mathrm{stage1}}, \mathcal{L}_{\mathrm{stage2}}$	First-stage and second-stage training objectives.
N, τ	Batch size and contrastive temperature.
$p_{\mathrm{src}}, p_{\mathrm{cfg}}$	Probabilities of source reconstruction and relation-condition dropout.
w	Relation-aware classifier-free guidance scale.

A PRELIMINARY

Flow Matching (Lipman et al., 2022) learns a continuous velocity field that transforms a Gaussian distribution into the data distribution. Given a clean latent representation z and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, we sample $t \sim \mathcal{U}(0, 1)$ and construct the interpolation path and target velocity as

$$z_t = (1 - t)z + t\epsilon, \qquad u_t = \epsilon - z. \tag{7}$$

A neural network $v_\theta(\cdot)$ with learnable parameters θ is trained to predict this velocity from z_t, the timestep t, and a conditioning context c. The optimization objective is

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{\epsilon,t}\left[\|v_\theta(z_t, t, c) - u_t\|_2^2\right]. \tag{8}$$

During inference, an ODE solver integrates the learned velocity field according to $\frac{d\hat{z}_t}{dt} = v_\theta(\hat{z}_t, t, c)$ from $t = 1$ to $t = 0$, where $\hat{z}_t$ denotes the latent state along the inference trajectory and $v_\theta(\hat{z}_t, t, c)$

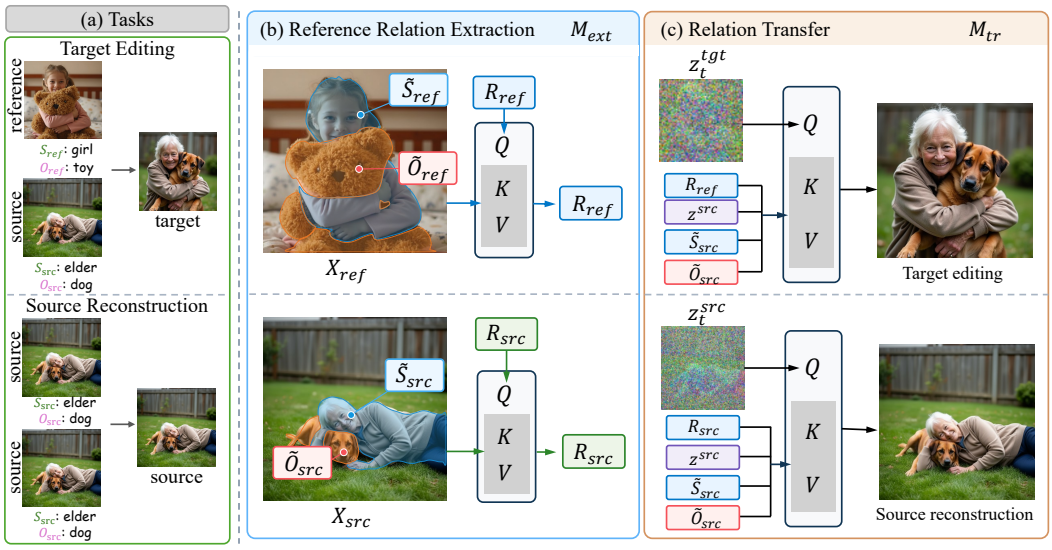

Figure 8: Visibility patterns in RTBA. (a) Target editing and source reconstruction tasks. (b) Relation extraction from the reference and source images. (c) Relation transfer for target editing and source reconstruction.

is the predicted velocity at timestep t under condition c. This process transforms an initial Gaussian noise sample $\hat{z}_1 \sim \mathcal{N}(0, I)$ into a clean data sample $\hat{z}_0$.

B Method Details

B.1 Details of RTBA

To provide a more intuitive explanation of the visibility patterns introduced by RTBA, Fig. 8 illustrates the effective information flow under $\mathbf{M}_{\text{ext}}$ and $\mathbf{M}_{\text{tr}}$.

As shown in Fig. 8(a), RACE performs two complementary tasks during training. In the target-editing task, the model extracts the relation from the reference image and transfers it to the designated subject–object pair in the source image to generate the target image. In the source-reconstruction task, the model extracts the original relation from the source image and reconstructs the source image itself, providing a source-side reconstruction trajectory under the original relation.

Fig. 8(b) illustrates relation extraction under $\mathbf{M}_{\text{ext}}$. For the reference image, $\widetilde{S}_{\text{ref}}$ and $\widetilde{O}_{\text{ref}}$ interact with the visual tokens X_{ref} to localize the designated subject and object, after which the learnable relation token R_{ref} aggregates the visual relation cues between them. The source image follows the same process, where $\widetilde{S}_{\text{src}}$ and $\widetilde{O}_{\text{src}}$ interact with X_{src}, and R_{src} aggregates the original relation of the designated source pair.

Fig. 8(c) illustrates relation transfer under $\mathbf{M}_{\text{tr}}$. In the target-editing branch, z_t^{tgt} uses the reference relation token R_{ref} together with the source-side conditions z^{src}, $\widetilde{S}_{\text{src}}$, and $\widetilde{O}_{\text{src}}$. Direct access from z_t^{tgt} to the reference visual representation X_{ref} is blocked, forcing reference relation information to pass through the limited-capacity relation token R_{ref}. In the source-reconstruction branch, z_t^{src} follows the same visibility pattern but uses R_{src} to preserve the original source relation and provide the source-side reconstruction.

B.2 Derivation of Semantic Shift Prediction

From Endpoint Semantic Shift to SSP. The ideal image-side semantic shift induced by relation editing is

$$\Delta^*_{\text{img}} = f_{\text{img}}(\hat{z}_0^{\text{tgt}}) - f_{\text{img}}(\hat{z}_0^{\text{src}}), \tag{9}$$

where $\hat{z}_0^{\mathrm{tgt}}$ and $\hat{z}_0^{\mathrm{src}}$ are generated by the target editing and source reconstruction branches, respectively, with $\hat{z}_t$ denoting the latent state at timestep t along the model-generated reverse denoising trajectory, and $f_{\mathrm{img}}(\cdot)$ denoting the image-side semantic encoder. We assume that, for each trajectory pair, the target editing and source reconstruction branches start from the same sampled Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ and are conditioned on the same source latent z^{src}. Conditioned on this shared noise, the model generates $\hat{z}_0^{\mathrm{src}}$ under the original source relation R_{src} and $\hat{z}_0^{\mathrm{tgt}}$ under the transferred relation R_{ref}. The two trajectories therefore satisfy

$$\begin{aligned} \hat{z}_0^{\mathrm{src}} &= \epsilon + \int_1^0 v_t^{\mathrm{src}}\, dt, \\ \hat{z}_0^{\mathrm{tgt}} &= \epsilon + \int_1^0 v_t^{\mathrm{tgt}}\, dt, \end{aligned} \tag{10}$$

where v_t^{src} and v_t^{tgt} denote the predicted velocities along the source reconstruction and target editing trajectories, respectively. Defining

$$\Delta v_t = v_t^{\mathrm{tgt}} - v_t^{\mathrm{src}}, \tag{11}$$

their endpoint difference becomes

$$\hat{z}_0^{\mathrm{tgt}} - \hat{z}_0^{\mathrm{src}} = \int_1^0 \Delta v_t\, dt, \tag{12}$$

We define the endpoint latent difference as

$$\Delta z_0 = \hat{z}_0^{\mathrm{tgt}} - \hat{z}_0^{\mathrm{src}}, \tag{13}$$

and, assuming that f_{img} is locally smooth around $\hat{z}_0^{\mathrm{src}}$, apply a first-order Taylor expansion to obtain

$$\begin{aligned} \Delta_{\mathrm{img}}^{*} &= f_{\mathrm{img}}(\hat{z}_0^{\mathrm{src}} + \Delta z_0) - f_{\mathrm{img}}(\hat{z}_0^{\mathrm{src}}) \\ &\approx J_{f_{\mathrm{img}}}(\hat{z}_0^{\mathrm{src}})\Delta z_0 \\ &= \int_1^0 J_{f_{\mathrm{img}}}(\hat{z}_0^{\mathrm{src}})\Delta v_t\, dt, \end{aligned} \tag{14}$$

where $J_{f_{\mathrm{img}}}(\cdot)$ denotes the Jacobian of f_{img}. As described in the main paper, the source reconstruction objective $\mathcal{L}_{\mathrm{src}}$ explicitly trains the source branch to recover the source latent under the original relation R_{src}, thereby providing a stable source-side trajectory. When the model is sufficiently well trained, its reconstruction endpoint closely approximates the source latent, i.e.,

$$\hat{z}_0^{\mathrm{src}} \approx z^{\mathrm{src}}. \tag{15}$$

We therefore use z^{src} in place of $\hat{z}_0^{\mathrm{src}}$, yielding

$$J_{f_{\mathrm{img}}}(\hat{z}_0^{\mathrm{src}}) \approx J_{f_{\mathrm{img}}}(z^{\mathrm{src}}). \tag{16}$$

Directly computing the Jacobian-vector product is impractical. We therefore introduce a Semantic Shift Predictor (SSP) to approximate the local semantic effect of Δv_t as

$$\Delta_{\mathrm{img}}^{(t)} = SSP\left(z^{\mathrm{src}}, \Delta v_t, t\right), \tag{17}$$

where $\Delta_{\mathrm{img}}^{(t)}$ denotes the image semantic shift predicted by SSP from the denoising dynamics at timestep t, and z^{src} provides the source identity, appearance, layout, background, and style context. For a fixed pair of trajectories initialized from the same Gaussian noise ϵ, absorbing the reverse-time sign into SSP, the endpoint semantic shift can be approximated as

$$\Delta_{\mathrm{img}}^{*} \approx \mathbb{E}_{t\sim\mathcal{U}(0,1)}\left[\Delta_{\mathrm{img}}^{(t)}\right]. \tag{18}$$

Timestep-Wise Semantic Shift Alignment. A direct objective would align the temporally aggregated semantic shift with the intended relation shift as

$$\mathcal{L}_{\mathrm{global}}^{\cos} = 1 - \cos\left(\mathbb{E}_{t\sim\mathcal{U}(0,1)}\left[\Delta_{\mathrm{img}}^{(t)}\right], \Delta_{\mathrm{text}}\right). \tag{19}$$

Algorithm 1 Training of RACE

Require: Dataset $\mathcal{D}$, Source-branch probability p_{src}, R-CFG dropout probability p_{cfg}, CSSA weight λ_{CSSA}, Warm-up steps N_1, CSSA fine-tuning steps N_2
Ensure: Model parameters θ, trainable LoRA parameters η of the latent-CLIP text encoder f_t, relation projection parameters ϕ, and SSP parameters ψ

Stage 1: Warm-up training without CSSA
for $s = 1$ to N_1 **do**
  Sample batch $(I_{\mathrm{ref}}, I_{\mathrm{src}}, I_{\mathrm{tgt}}, S_{\mathrm{ref}}, O_{\mathrm{ref}}, S_{\mathrm{src}}, O_{\mathrm{src}}, r_{\mathrm{src}}, r_{\mathrm{tgt}})$ from $\mathcal{D}$
  Obtain noisy latents z_t^{src} and z_t^{tgt}, and analytic velocity targets u_t^{src} and u_t^{tgt} via flow matching (Eq. 7 in Appendix A).
  if reconstruct original branch with p_{src} **then**
    Update θ using the source reconstruction objective $\mathcal{L}_{\mathrm{src}}$
  else
    Drop R_{ref} with probability p_{cfg}
    Update θ using the target editing objective $\mathcal{L}_{\mathrm{tgt}}$
  end if
end for

Stage 2: Joint training with CSSA
for $s = 1$ to N_2 **do**
  Repeat Stage 1 preprocessing to sample batch, encode inputs, and construct noisy latents.
  if reconstruct original branch with p_{src} **then**
    Update θ using the source reconstruction objective $\mathcal{L}_{\mathrm{src}}$
  else
    Drop R_{ref} with probability p_{cfg}
    Update θ, η, ϕ, and ψ using the target editing objective together with the CSSA objective
  end if
end for

return θ, η, ϕ, ψ

However, this objective only constrains the accumulated endpoint shift. Relation-irrelevant semantic shifts at different timesteps may cancel after temporal aggregation. For example, consider two equally weighted timesteps with

$$\Delta_{\mathrm{img}}^{(t_1)} = \Delta_{\mathrm{text}} + \delta_{\mathrm{irr}}, \qquad \Delta_{\mathrm{img}}^{(t_2)} = \Delta_{\mathrm{text}} - \delta_{\mathrm{irr}}, \tag{20}$$

where δ_{irr} represents a relation-irrelevant semantic shift in identity, appearance, background, or style. Their average still equals Δ_{text}, although both timesteps contain undesired semantic shifts.

We therefore apply the constraint before temporal aggregation. A timestep-wise cosine objective can be written as

$$\mathcal{L}_{\mathrm{step}}^{\cos} = \mathbb{E}_{t\sim\mathcal{U}(0,1)}\left[1 - \cos\left(\Delta_{\mathrm{img}}^{(t)}, \Delta_{\mathrm{text}}\right)\right]. \tag{21}$$

This timestep-wise objective provides a finer constraint by penalizing relation-irrelevant semantic shifts as they emerge, rather than only evaluating their accumulated endpoint effect. In practice, rather than directly optimizing this positive-only cosine objective, we use the contrastive CSSA objective defined in the main paper, which additionally introduces in-batch negative relation shifts.

B.3 SSP Architecture Details

SSP consists of two convolutional encoders, a time-embedding module, a source-conditioned gating and fusion module, and an MLP prediction head. The two encoders extract features from the source latent and the velocity difference, respectively, with timestep information injected via FiLM (Perez et al., 2018). Channel-wise gates derived from the source features modulate the velocity-difference features before fusion. Finally, the globally pooled source features, pre-gating velocity-difference features, and fused features are concatenated with the time embedding and mapped by the prediction head to an L2-normalized semantic-shift direction for CSSA.

Algorithm 2 Inference of RACE

Require: Reference image I_{ref}, source image I_{src}, object texts S_{ref}, O_{ref}, S_{src}, O_{src}
Require: R-CFG guidance scale w, denoising steps T
Require: Trained model $v_\theta(\cdot)$, learnable relation token R_{ref}
Ensure: Edited target image $\hat{I}_{\mathrm{tgt}}$
1: Set $t_k = k/T$ for $k = 0, \ldots, T$ and initialize $\hat{z}^{\mathrm{tgt}}_{t_T} \sim \mathcal{N}(0, I)$
2: **for** $k = T$ down to 1 **do**
3: Obtain $v^{\mathrm{tgt}}_{t_k}$ using Eq. 6.
4: Update $\hat{z}^{\mathrm{tgt}}_{t_k}$ to $\hat{z}^{\mathrm{tgt}}_{t_{k-1}}$ following flow matching as described in Appendix A.
5: **end for**
6: Obtain $\hat{I}_{\mathrm{tgt}}$ from $\hat{z}^{\mathrm{tgt}}_{t_0}$ via VAE decoding.
7:
8: **return** $\hat{I}_{\mathrm{tgt}}$

B.4 Training and Inference Details

We adopt a two-stage training strategy. In the first stage, we optimize the Bagel editing backbone for 6,650 training steps without activating CSSA. For each training step, we perform either the target image generation task or the source reconstruction task. For target editing, the model is optimized with the flow-matching objective

$$\mathcal{L}_{\mathrm{tgt}} = \mathbb{E}_{t\sim\mathcal{U}(0,1),\,\epsilon\sim\mathcal{N}(0,I)}\left[\left\| v_\theta\left(z^{\mathrm{tgt}}_t, t, c^{\mathrm{tgt}}\right) - u^{\mathrm{tgt}}_t \right\|_2^2\right], \tag{22}$$

where v_θ and u^{tgt}_t denote the predicted and ground-truth velocities, respectively, and $c^{\mathrm{tgt}} = [R_{\mathrm{ref}}, \widetilde{S}_{\mathrm{src}}, \widetilde{O}_{\mathrm{src}}, z^{\mathrm{src}}]$ denotes the target-editing condition. With probability p_{src}, we instead perform source reconstruction and optimization with

$$\mathcal{L}_{\mathrm{src}} = \mathbb{E}_{t\sim\mathcal{U}(0,1),\,\epsilon\sim\mathcal{N}(0,I)}\left[\left\| v_\theta\left(z^{\mathrm{src}}_t, t, c^{\mathrm{src}}\right) - u^{\mathrm{src}}_t \right\|_2^2\right], \tag{23}$$

where $c^{\mathrm{src}} = [R_{\mathrm{src}}, \widetilde{S}_{\mathrm{src}}, \widetilde{O}_{\mathrm{src}}, z^{\mathrm{src}}]$. The first-stage objective is therefore

$$\mathcal{L}_{\mathrm{stage1}} = \begin{cases} \mathcal{L}_{\mathrm{src}}, & \text{with probability } p_{\mathrm{src}}, \\ \mathcal{L}_{\mathrm{tgt}}, & \text{with probability } 1 - p_{\mathrm{src}}. \end{cases} \tag{24}$$

In the second stage, we activate CSSA and continue training for an additional 1,200 steps, jointly fine-tuning the Bagel backbone, latent-CLIP text encoder f_t, relation projection network g_ϕ, and SSP. The target-editing and source-reconstruction branches share the same sampled timestep t and Gaussian noise ϵ, ensuring that their trajectory-level velocity difference is computed under matched perturbations. CSSA is applied to the target-editing branch to constrain the editing-induced semantic change toward the intended relation shift, while the source reconstruction branch remains optimized with $\mathcal{L}_{\mathrm{src}}$. The training objective becomes

$$\mathcal{L}_{\mathrm{stage2}} = \begin{cases} \mathcal{L}_{\mathrm{src}}, & \text{with probability } p_{\mathrm{src}}, \\ \mathcal{L}_{\mathrm{tgt}} + \lambda_{\mathrm{CSSA}}\mathcal{L}_{\mathrm{CSSA}}, & \text{with probability } 1 - p_{\mathrm{src}}. \end{cases} \tag{25}$$

We activate CSSA only after the first-stage warm-up because early denoising predictions and source reconstruction trajectories are insufficiently stable for reliable semantic-shift estimation. Applying CSSA too early would therefore introduce noisy semantic supervision and potentially destabilize training. Detailed training and inference algorithms are provided in Algorithms 1 and 2.

B.5 Editability-Identity Score and User Study

Editability-Identity Score (EI). Inspired by OneHOI (Hoe et al., 2026), we define the editability-identity score (EI) to jointly evaluate relation editability and identity preservation. For each sample,

Table 4: Comparison of training and inference efficiency.

Method	Training Time (h)	Training Steps	Inference Time (s/image)	Inference Memory (GB)
FLUX.1-Kontext (arXiv'25) (Labs et al., 2025)	99.63	7,850	62.15	37.36
OneHOI (CVPR'26) (Hoe et al., 2026)	92.61	7,850	54.03	42.33
Baseline (Bagel, arXiv'25) (Deng et al., 2025)	64.89	7,850	35.33	29.71
Ours	78.55	$6{,}650 + 1{,}200$	35.08	29.77

DINO-I is computed as the cosine similarity between the last-layer DINOv2 (Oquab et al., 2023) features of the source and generated images. EI is computed as the harmonic mean of Relation Consistency and DINO-I:

$$\mathrm{EI} = \frac{2 \times \text{Relation Consistency} \times \text{DINO-I}}{\text{Relation Consistency} + \text{DINO-I}}. \tag{26}$$

A higher EI indicates a better balance between relation transfer and source-image preservation.

User Study. We invited 24 experts to participate in the user study, with all participants evaluating the same set of samples. The study covered 13 methods across 8 distinct reference–source image pairs. To reduce the ranking burden, the 13 methods were distributed across evaluation groups, with six methods presented in each group. The display order was randomized and kept identical across participants. Participants independently ranked the results along three dimensions: (1) relation consistency, (2) editing consistency, and (3) image quality, with no ties allowed. For each dimension, ranks from 1 to 6 were converted into normalized scores from 100 to 0, with higher scores indicating stronger preference. The final score of each method was obtained by averaging its scores across all evaluations and participants.

C Training and Inference Efficiency

As shown in Table 4, we compare the training and inference efficiency of our method with FLUX.1-Kontext, OneHOI, and the Bagel-based baseline. Our method reduces training time by 21.2% and 15.2% compared with FLUX.1-Kontext and OneHOI, respectively, mainly benefiting from the causal attention mechanism adopted by Bagel, which avoids unnecessary token interactions and reduces multimodal attention cost. Compared with the baseline, our method introduces additional training overhead because the second stage jointly optimizes the editing backbone, f_t, g_ϕ, and the Semantic Shift Predictor (SSP). Nevertheless, these additional components introduce almost no inference overhead. Compared with the Bagel-based baseline, our method maintains nearly identical inference efficiency and peak memory usage, while reducing inference time by 43.6% and 35.1% compared with FLUX.1-Kontext and OneHOI, respectively.

D Additional Experiments and Analysis

We provide additional experiments and analyses to further evaluate the effectiveness and generalization of RACE.

D.1 Training Configuration

For CSSA, we fine-tune the last two latent-CLIP layers using LoRA (Hu et al., 2021) of rank 16 and use an 8-layer MLP for the relation projection network g_ϕ. The editing backbone is fine-tuned using LoRA of rank 32. We use AdamW (Kingma & Ba, 2014; Loshchilov & Hutter, 2017) with a learning rate of 1×10^{-4}.

D.2 Hyperparameter experiments

As shown in Table 5, we analyze the CSSA fine-tuning steps, R-CFG scale w, and λ_{CSSA}. Increasing CSSA training from 600 to 1,200 steps improves source-image preservation, while extending it to 2,100 steps degrades both relation consistency and source-image preservation. This may occur because prolonged training encourages SSP to shortcut the prediction by directly extracting relation

Table 5: Hyperparameter experiments on RTBench. Ours uses $N_2 = 1{,}200$, $w = 3.5$, and $\lambda_{\text{CSSA}} = 2.0$.

Parameter	Value	Our setting	relation consistency↑	EI↑	HPSv2↑	Image Reward↑	PickScore↑
N_2	600	1,200	**0.3182**	<u>0.3821</u>	<u>0.2784</u>	<u>0.9413</u>	<u>21.9392</u>
N_2	2,100	1,200	0.2380	0.3045	0.2703	0.6932	21.7103
w	2	3.5	0.2055	0.2825	0.2721	0.6897	21.7884
w	5	3.5	0.2891	0.3624	0.2775	0.9085	21.9068
λ_{CSSA}	0.5	2.0	0.2672	0.3377	0.2700	0.7223	21.7593
λ_{CSSA}	3.5	2.0	0.2900	0.3530	0.2698	0.7045	21.8534
Ours	$N_2 = 1{,}200$	$w = 3.5, \lambda_{\text{CSSA}} = 2.0$	<u>0.2985</u>	**0.3851**	**0.2913**	**1.2699**	**22.4189**

Table 6: Comparison with alternative sparse-attention designs on RTBench.

Method	relation consistency↑	EI↑	HPSv2↑	Image Reward↑	PickScore↑
SIGMA-style Group-Scoped Mask (Zhang et al., 2026)	0.1656	0.2398	0.2558	0.2512	21.2910
OrionEdit-style Reverse-Causal Mask (Jiang et al., 2026)	0.1353	0.2097	0.2529	0.2234	21.2264
RTBA (Ours)	**0.2985**	**0.3851**	**0.2913**	**1.2699**	**22.4189**

semantic shifts from Δv_t, rather than modeling the complete image semantic shift, thereby weakening its ability to detect and constrain appearance variations. For R-CFG, $w = 2$ provides insufficient relation guidance, often resulting in incomplete edits or retention of the source relation. Increasing the scale to $w = 3.5$ strengthens relation control while maintaining source-image consistency, whereas $w = 5$ over-amplifies the relation-conditioned direction, leading to excessive structural deformation, identity or appearance shifts, and degraded local details. For CSSA, $\lambda_{\text{CSSA}} = 2$ outperforms $\lambda_{\text{CSSA}} = 0.5$ because stronger semantic alignment not only suppresses relation-irrelevant semantic shifts but also concentrates editing-induced shifts on the intended relation direction. However, $\lambda_{\text{CSSA}} = 3.5$ overemphasizes semantic alignment relative to the flow-matching objective, leading to degraded source preservation and generation quality.

D.3 Comparison with Alternative Sparse Attention Designs

We further compare RTBA with two representative sparse-attention designs, including the group-scoped mask in SIGMA (Zhang et al., 2026) and the reverse-causal mask in OrionEdit (Jiang et al., 2026), while keeping the same Bagel backbone and training settings. As shown in Table 6, RTBA achieves consistently better results across relation consistency, source-image preservation, and perceptual quality metrics, indicating that its relation-bottleneck design more effectively separates transferable relation cues from reference-specific appearance.

D.4 Fair comparison with reference-conditioned baselines

To isolate the effect of reference-image access and training data, we adapt OneHOI and FLUX Kontext by directly concatenating the reference image with the noisy input as an additional visual condition. Both adapted baselines are trained on RTD-35K for the same 7,850 steps using the same training configuration as RACE. As shown in Table 1, RACE consistently outperforms both reference-conditioned baselines across all metrics. These results demonstrate that the improvements of RACE arise not merely from reference-image access or RTD-35K training, but from its structured relation isolation and semantic-shift regularization.

D.5 Generalization Across Different Base Models

To evaluate whether our method is tied to a specific backbone, we further integrate RTBA and CSSA into different image editing models, including FluxKontext, Qwen Image Edit 2511, and Bagel. As



Figure 9: Generalization across different base models. Integrating RTBA and CSSA into FluxKontext, Qwen Image Edit 2511, and Bagel consistently improves relation editing and source-image preservation.

Table 7: Quantitative comparison on RTBench.

Method	relation consistency↑	EI↑	HPSv2↑	Image Reward↑	PickScore↑
MSDiffusion (ICLR'25) (Wang et al., 2025)	0.1754	0.2550	0.2808	0.6704	21.6361
DreamRelation (CVPR'25) (Shi et al., 2025)	0.1933	0.2766	0.2841	0.8991	21.8448
DyRef (ECCV'26) (Huang et al., 2026)	**0.3443**	0.3074	0.2389	-0.4566	20.1848
FluxKontext (arXiv'25) (Labs et al., 2025)	0.1045	0.1926	0.2858	0.6682	22.0042
FluxKontext (w/ ref.) (Labs et al., 2025)	0.2828	0.3527	0.2568	0.3691	21.2824
OmniGen2 (arXiv'25) (Wu et al., 2025b)	0.1129	0.2022	0.2782	0.6375	21.8879
qwen-image-edit (arXiv'25) (Wu et al., 2025a)	0.1130	0.1879	0.2631	-0.0251	21.2682
HOIEdit (TVC'25) (Xu et al., 2025a)	0.0818	0.1601	0.2293	-0.5048	20.2097
InteractEdit (arXiv'25) (Hoe et al., 2025)	0.1842	0.2374	0.2829	0.6948	21.8931
OneHOI (CVPR'26) (Hoe et al., 2026)	0.1632	0.2456	0.2912	0.9340	22.0644
OneHOI (w/ ref.) (Hoe et al., 2026)	0.2181	0.2891	0.2861	1.0794	22.1754
nano banana (commercial) (Google, 2025)	0.2617	0.2817	0.2688	0.4586	21.4094
Ours	0.2985	**0.3851**	**0.2913**	**1.2699**	**22.4189**

shown in Fig. 9, incorporating our modules consistently improves relation editing across different base models. These results demonstrate that our method is not specific to Bagel and can generalize to different image editing architectures.

D.6 Detailed Quantitative Results on RTBench

We provide the complete quantitative comparison on RTBench in Table 7. RACE achieves the highest EI, HPSv2, ImageReward, and PickScore, demonstrating a strong balance between relation transfer, source-image preservation, and generation quality.

D.7 Additional Results on Generalization and Extensibility

Figures 10 and 11 present additional qualitative comparisons with open-source and commercial image editing models, respectively. The results show that RACE achieves state-of-the-art performance

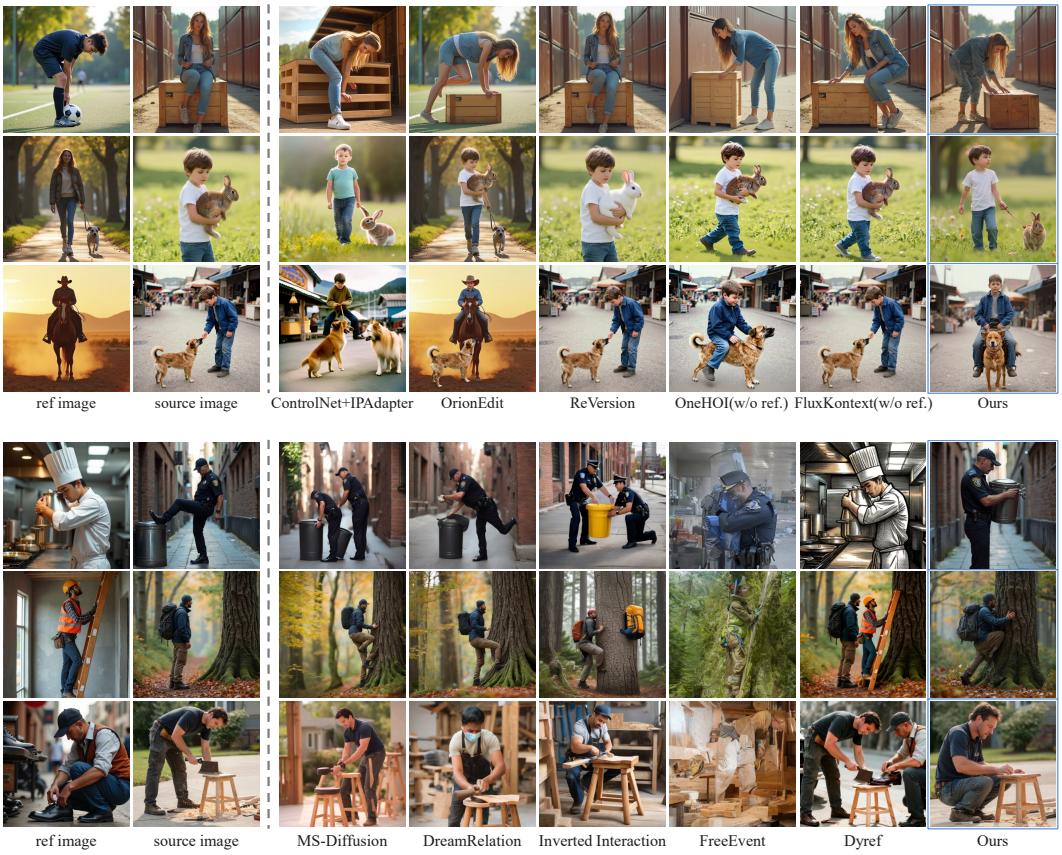

Figure 10: Qualitative comparison with open-source methods (Zhang et al., 2023c; Ye et al., 2023; Jiang et al., 2026; Huang et al., 2024; Hoe et al., 2026; Labs et al., 2025; Wang et al., 2025; Shi et al., 2025; Ge et al., 2024; Wang et al., 2024; Huang et al., 2026).

among open-source methods and remains competitive with strong closed-source commercial models.

Fig. 12 presents diverse reference–source combinations, illustrating the robustness of RACE across different reference and source images. Although RTD-35K contains only human-centered relations, Fig. 13 shows that RACE can generalize relation transfer to diverse object–object interactions. Fig. 15 further shows that RACE supports multi-group relation transfer, jointly transferring multiple relations from a single reference image to the corresponding object groups in the source image. Fig. 14 demonstrates the extensibility of RACE to skeleton- and layout-guided editing.

To study generalization to unseen relations, we deliberately remove four relation categories—*shake hands*, *fist bump*, *kiss*, and *bow to*—from the training data and treat them as held-out relations. Fig. 16 shows qualitative results on these categories, demonstrating that RACE can transfer relations that are never observed during training.

E Failure Cases and Limitations

As shown in Fig. 21, the failure cases of RACE mainly fall into two categories. First, although the global relation can often be transferred correctly, fine-grained interaction regions may still contain local artifacts, as illustrated in Fig. 21(a). Second, relation transfer becomes substantially more difficult when the target relation requires drastic structural modifications. In such cases, the model must jointly modify the human pose, object location, and overall scene layout, which may result in relation transfer failure, as shown in Fig. 21(b).



Figure 11: Qualitative comparison with commercial image editing models (Reve Team, 2026; Google, 2025; 2026; OpenAI, 2026). RACE demonstrates highly competitive relation transfer performance while preserving the identity characteristics and appearance of source objects.



Figure 12: Diverse reference–source combinations for relation transfer.

These limitations suggest two possible directions for future improvement. For interaction-region artifacts, inspired by the masked attention mechanism in DreamStory (He et al., 2025), region-aware attention constraints could be introduced so that generated object regions attend only to the corresponding source-object regions, while background regions attend to the source background. For drastic structural modifications, an adaptive editing mask could be introduced to relax source-preservation constraints around the interacting subject and object, while keeping unrelated background regions fixed. This would provide the model with more flexibility to modify pose, object position, and local layout when large structural changes are required.

F Details of RTD-35K Construction

The data construction pipeline is illustrated in Fig. 19. We first collect candidate relation categories covering diverse scenarios, including daily activities, work, sports, social interactions, transportation, and object manipulation. For each relation, we instantiate diverse subject–object combinations and scene contexts, and use GPT-5-mini to generate structured image descriptions, with the detailed prompts shown in Fig. 18. Based on these descriptions, we synthesize the reference image with FLUX.1-dev, generate the target image using the extracted reference skeleton with a ControlNet-

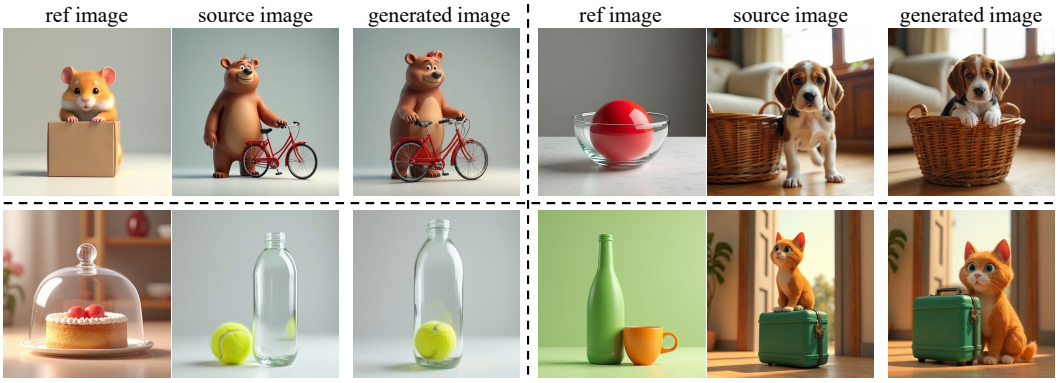

Figure 13: Object–object relation transfer examples. Although RACE is trained on human-centered relations, it can generalize relation transfer to diverse object pairs.

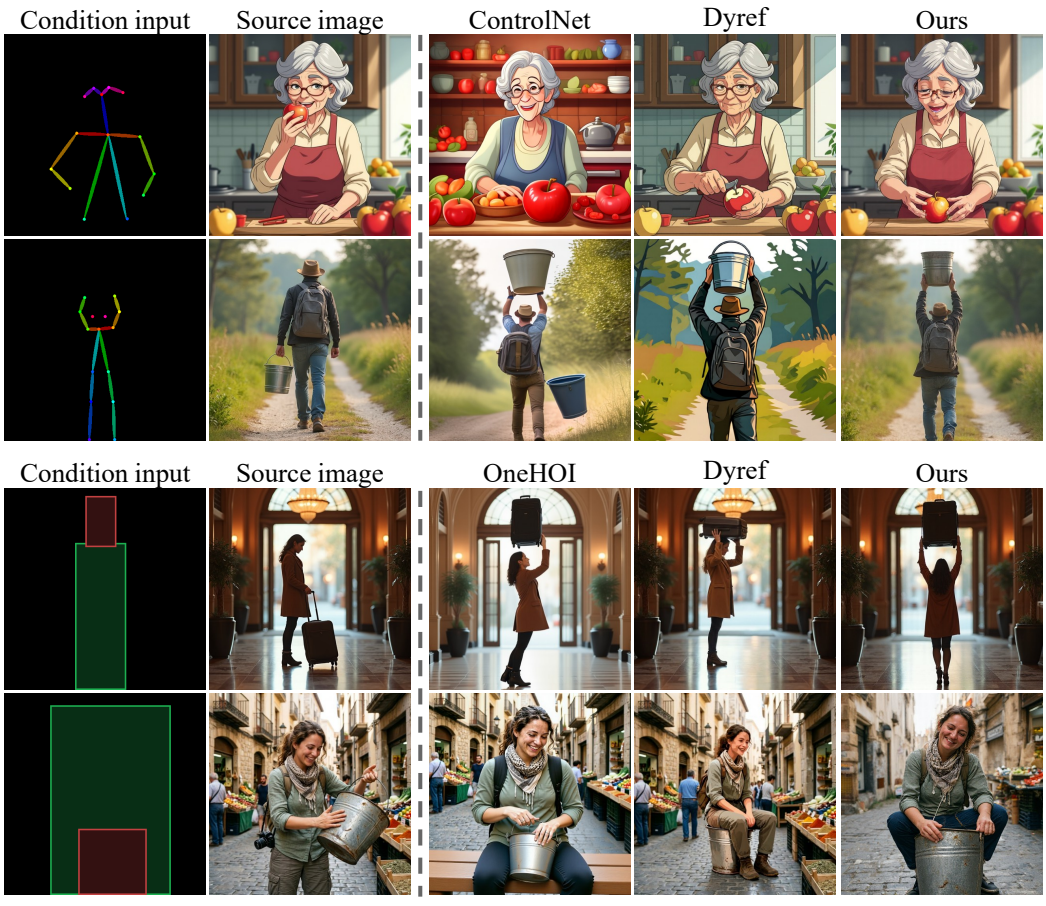

Figure 14: Extending RACE to conventional skeleton- and layout-guided editing.

based pipeline, and construct the source image by editing the target with Qwen-Image-Edit to introduce a different relation while preserving appearance. After generation, we use GPT-5-mini to filter the triplets. We retain samples with consistent reference–target relations and source–target appearance, while requiring different source and target relations. Approximately 80% of generated samples are filtered out during this process.

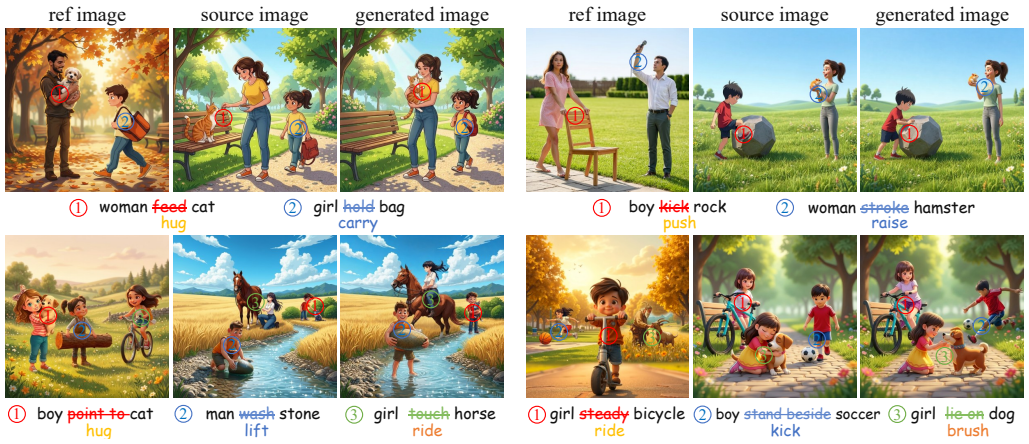

Figure 15: Multi-group relation transfer with RACE.

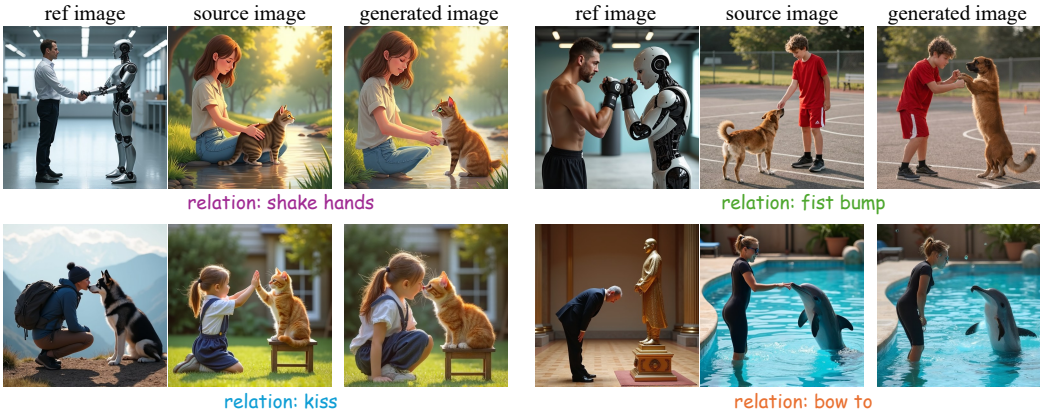

Figure 16: Qualitative results on held-out relation categories. We deliberately exclude *shake hands*, *fist bump*, *kiss*, and *bow to* from the training data. Despite never observing these relations during training, RACE can still transfer their relation semantics from the reference image to the source image, demonstrating its generalization to unseen relations.

In total, RTD-35K contains 35K reference–source–target triplets spanning 235 relation categories. Fig. 20 and Fig. 17 visualize the relation categories and their frequency distribution, respectively. Representative examples of the *carry*, *hold*, *lift*, and *push* relations are shown in Figs. 22, 23, 24, and 25, respectively.

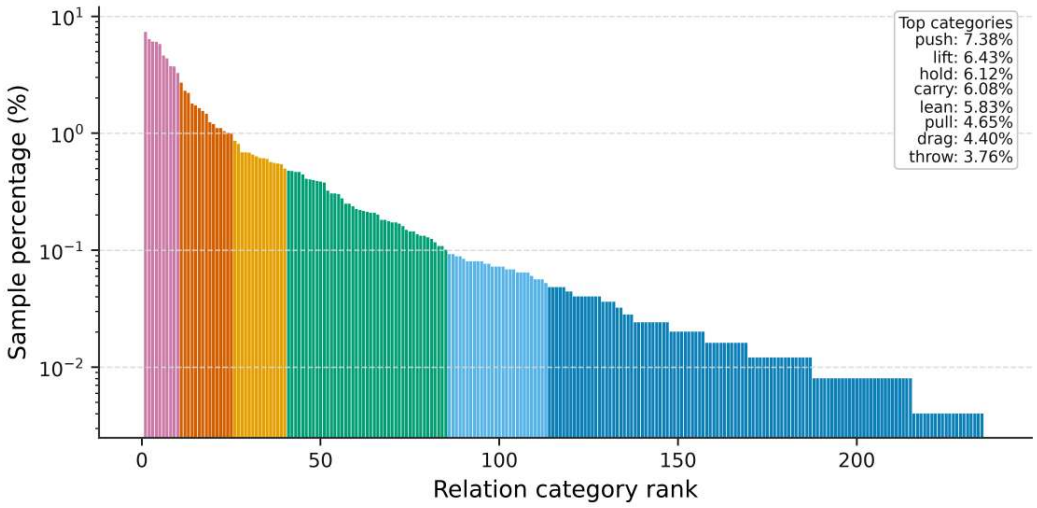

Figure 17: Distribution of the 235 relation categories in RTD-35K.

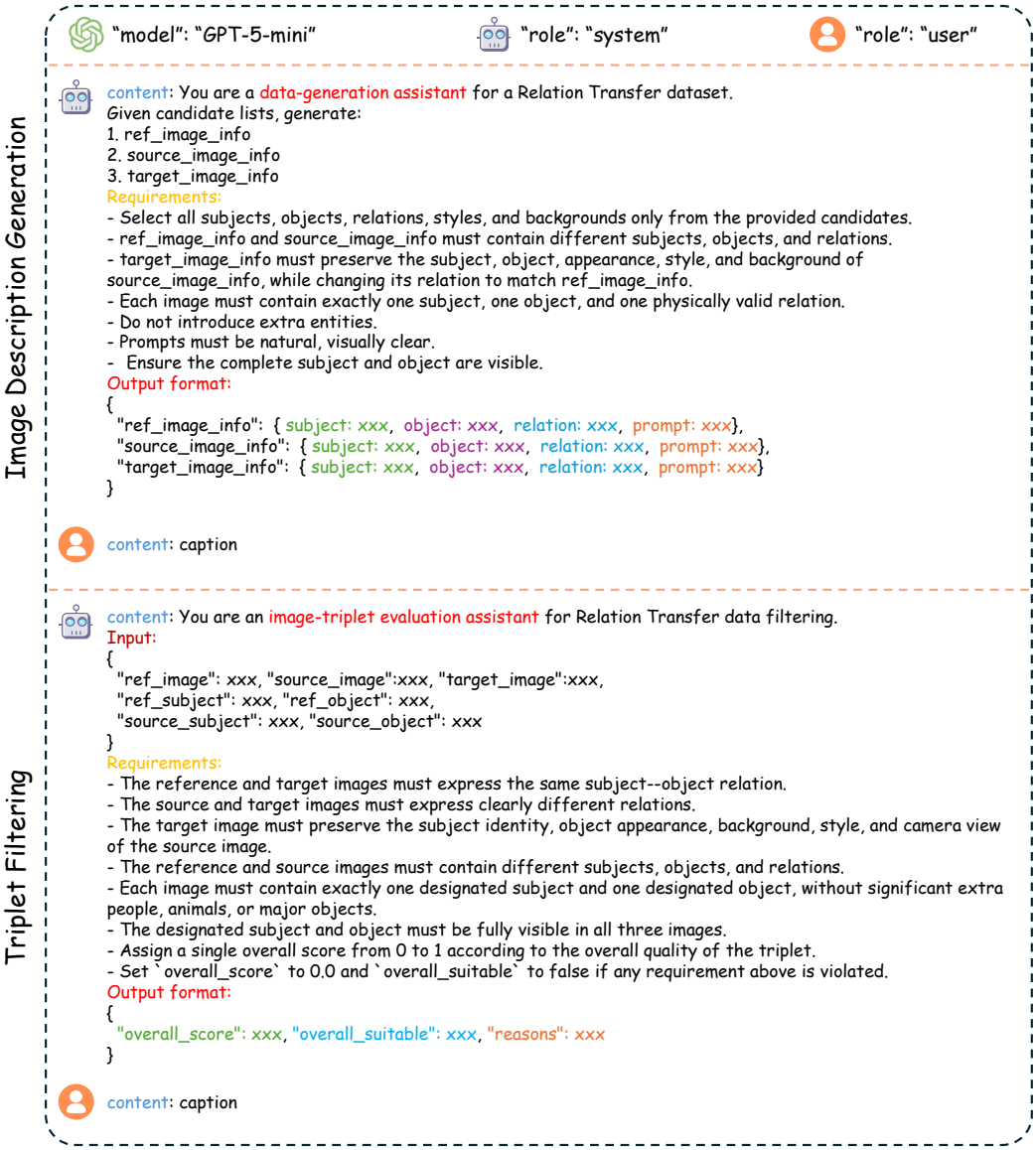

Figure 18: GPT prompt template for constructing RTD-35K.

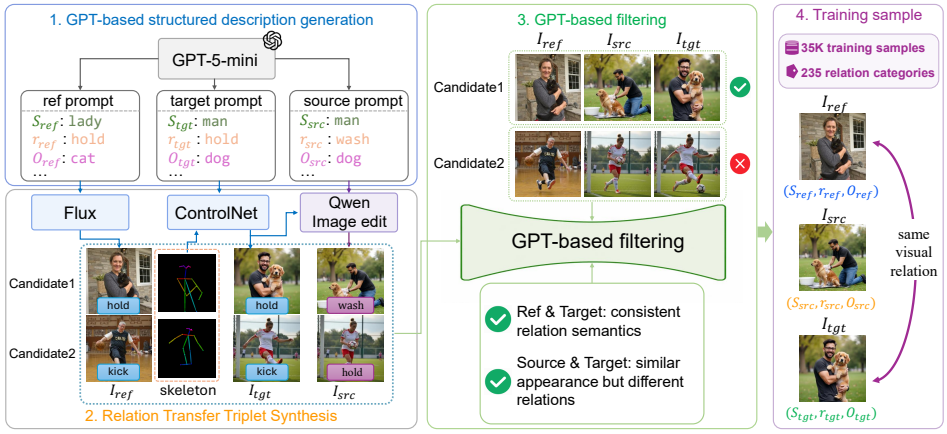

Figure 19: Pipeline for constructing the paired 35K-scale Relation Transfer Dataset (RTD-35K).

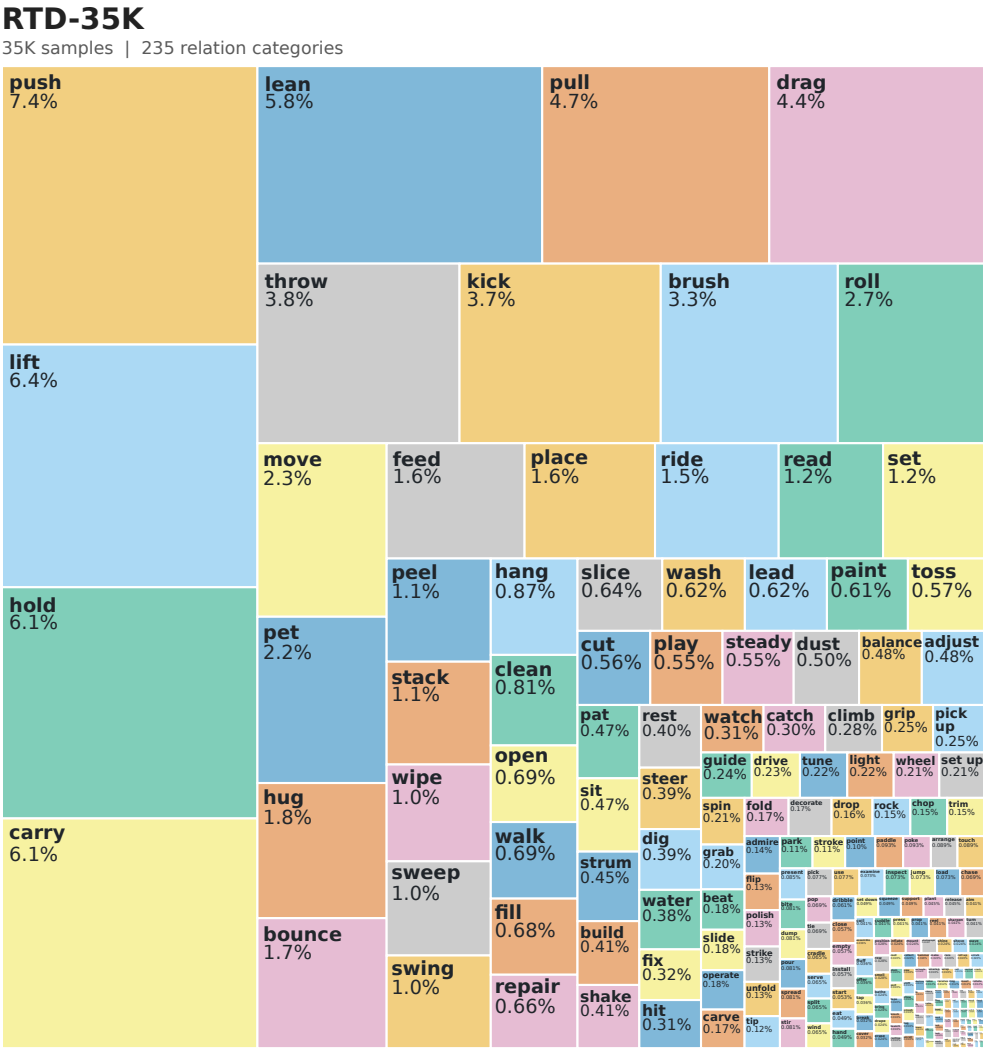

Figure 20: Treemap visualization of the 235 relation categories in RTD-35K.

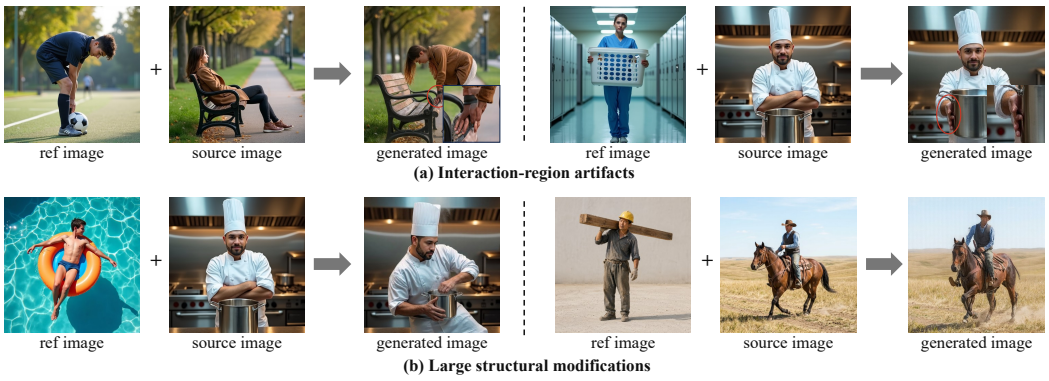

Figure 21: Representative failure cases of RACE. (a) Artifacts in interaction regions. (b) Relation transfer failures under drastic structural modifications.

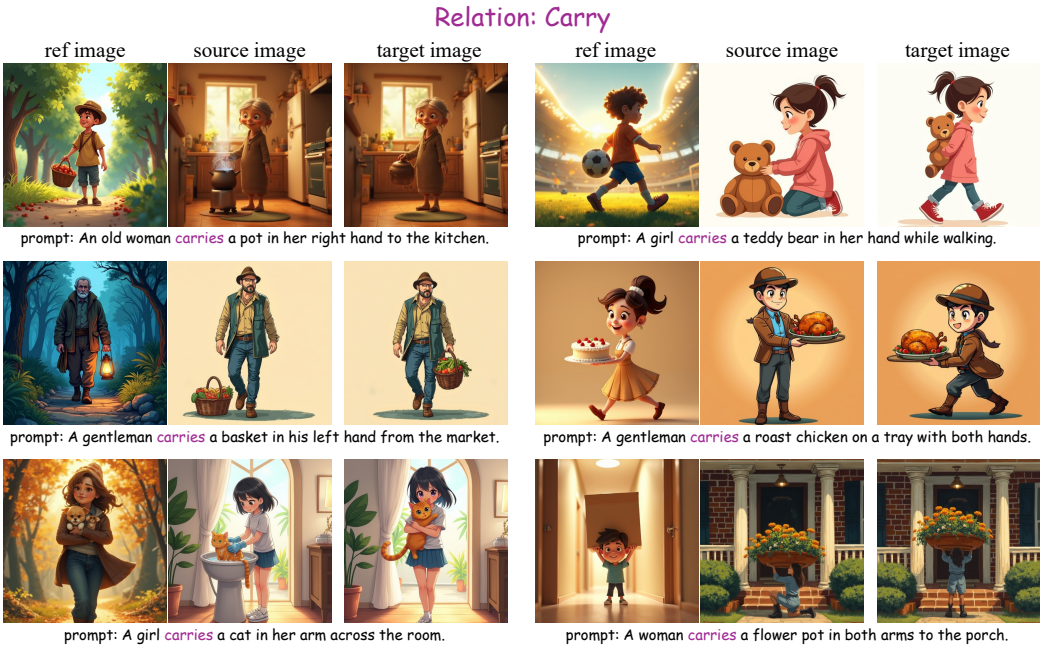

Figure 22: Examples for the *carry* relation in RTD-35K.

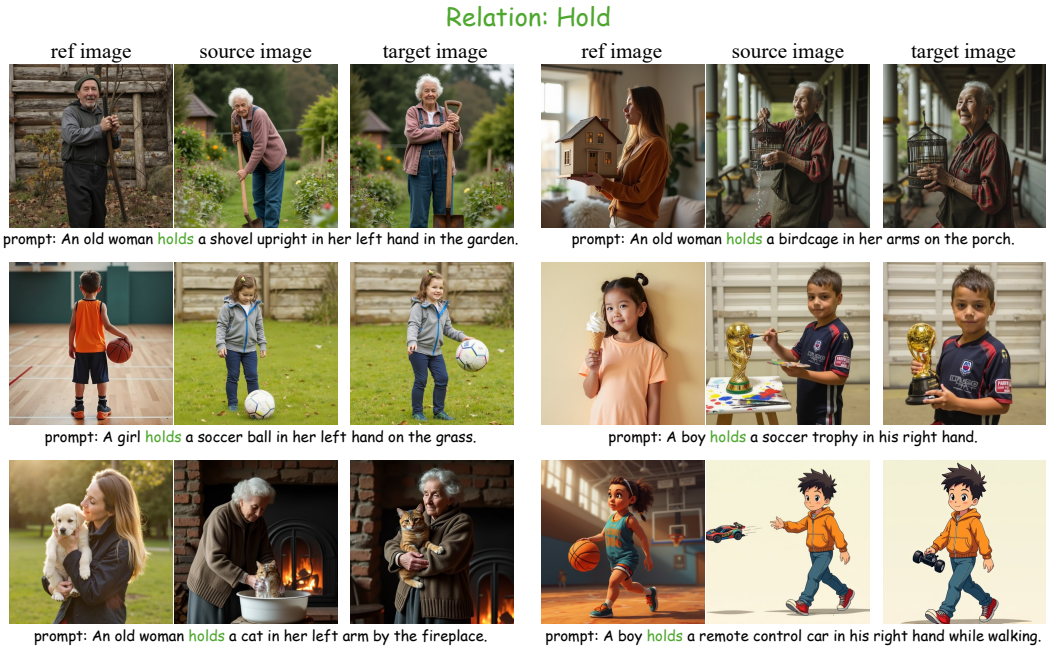

Figure 23: Examples for the *hold* relation in RTD-35K.

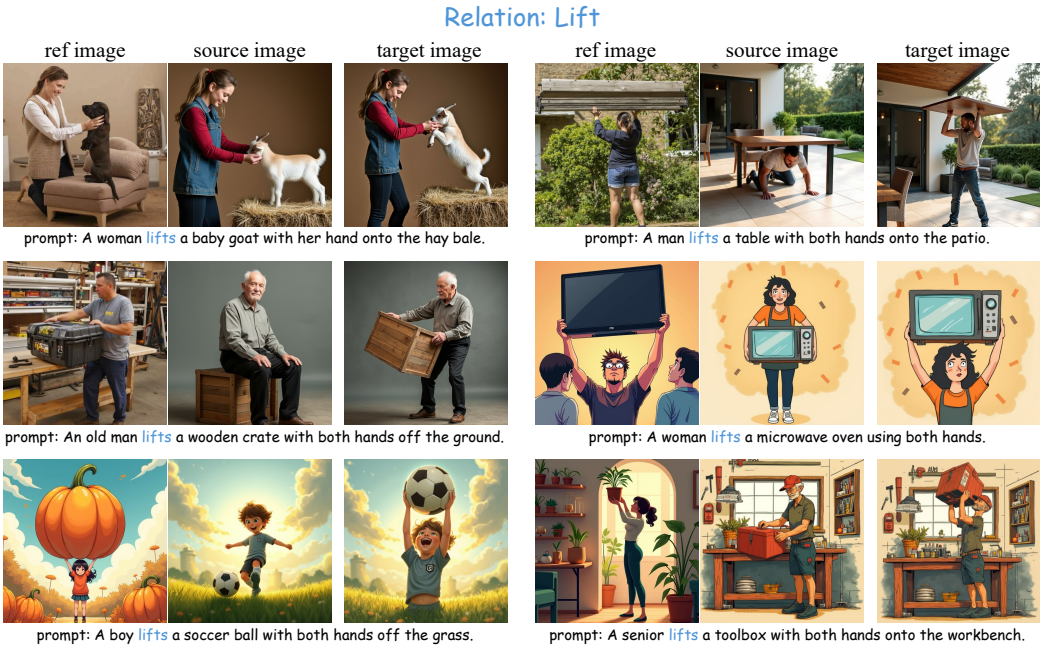

Figure 24: Examples for the *lift* relation in RTD-35K.



Figure 25: Examples for the *push* relation in RTD-35K.