Neurocomputing 704 (2026) 134869

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Reference-free human-object interaction editing

Jiun Tian Hoe [a], Weipeng Hu [a,b,*], Wei Zhou [b], Chao Xie [c], Ziwei Wang [a], Xudong Jiang [a], Yap-Peng Tan [a,e], Chee Seng Chan [d,e,**]

[a] School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Ave, 639798, Singapore, Singapore
[b] School of Intelligent Systems Engineering, Sun Yat-sen University, 518107, Shenzhen Campus, Guangdong Province, China
[c] College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing, 210037, Jiangsu Province, China
[d] Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, 50603, Malaysia
[e] College of Engineering and Computer Science, VinUniversity, Hanoi, 12426, Viet Nam

HIGHLIGHTS

- We define the reference-free HOI editing task and introduce the IEBench benchmark.
- Disassembling HOI scenes disentangles complex subject-object relationships.
- Selective Inversion and SeRA enable non-rigid HOI edits while preserving identity.
- We achieve the best balance between HOI editability and identity consistency.

ARTICLE INFO

Communicated by W. Zhou

Keywords:
Human-object interaction
Image editing
Diffusion models
Generative models
Deep learning
Computer vision

ABSTRACT

This paper presents InteractEdit, a novel framework for reference-free Human-Object Interaction (HOI) editing that tackles the challenging task of transforming an existing interaction in an image into a new, desired interaction while preserving the identities of the subject and object. Unlike prior image editing tasks such as attribute manipulation, object replacement or style transfer, HOI editing involves complex spatial, contextual, and relational dependencies inherent in HOI. Existing methods often overfit to the source image structure, limiting adaptability to the substantial structural modifications demanded by the new interactions. To address this, InteractEdit disassembles each scene into subject, object, and background components to disentangle intricate HOI relationships, and introduces a selective inversion strategy combined with Selective-Rank Adaptation (SeRA) to leverage pretrained interaction priors while learning visual identity from the source image. This enables a balanced trade-off between interaction editing and identity preservation. We also introduce IEBench, a new benchmark for HOI editing, and a new metric that jointly evaluates the trade-off between successful interaction editing and identity preservation. Extensive experiments show that InteractEdit outperforms 23 existing methods, providing a strong baseline for future HOI editing research. Code and the dataset: https://jiuntian.github.io/InteractEdit/.

1. Introduction

Human–Object Interaction (HOI) understanding plays a key role in computer vision, supporting a wide range of applications from activity recognition [1] and augmented-reality overlays [2] to robotics [3]. While HOI *recognition* and *generation* have been well studied [4–7], the task of *editing an existing interaction within a single image* remains largely unexplored despite its clear commercial potential. In industries such as film and media, modifying a character's action, for example from *riding a horse* to *kissing a horse* can save an entire re-shoot. Similarly, in e-commerce and advertising, updating a product visual, e.g. changing an athlete from *holding* to *drinking* a sports drink, can significantly reduce production time and expenses. These scenarios demand a careful

* Corresponding author at: School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Ave, 639798, Singapore.
** Corresponding author at: Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, 50603, Malaysia.
Email addresses: jiuntian001@e.ntu.edu.sg (J.T. Hoe), huwp7@mail.sysu.edu.cn (W. Hu), zhouw75@mail3.sysu.edu.cn (W. Zhou), chaoxie@njfu.edu.cn (C. Xie), ziwei.wang@ntu.edu.sg (Z. Wang), exdjiang@ntu.edu.sg (X. Jiang), eyptan@ntu.edu.sg (Y.-P. Tan), yp.t@vinuni.edu.vn (Y.-P. Tan), cs.chan@um.edu.my (C.S. Chan).

https://doi.org/10.1016/j.neucom.2026.134869
Received 23 March 2026; Received in revised form 3 August 2026; Accepted 18 August 2026
Available online 19 August 2026
0925-2312/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

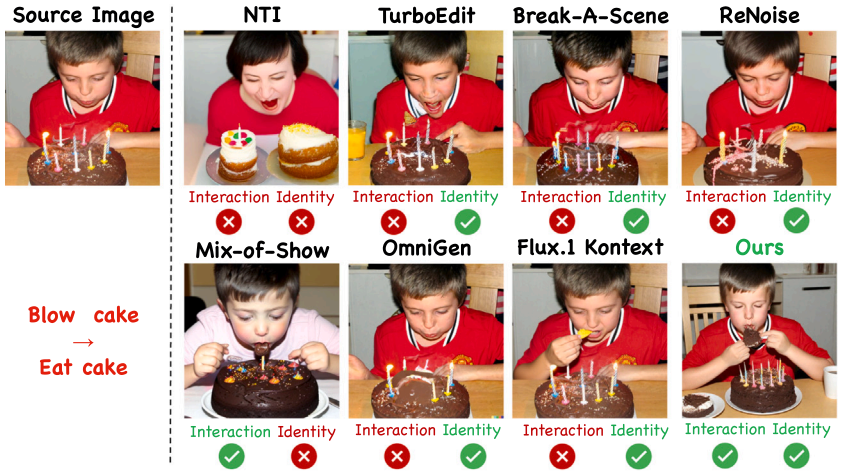


Fig. 1. Sample results of editing Human-Object interaction in the source image (left). Existing methods overly preserve the structure, making interaction edits ineffective.

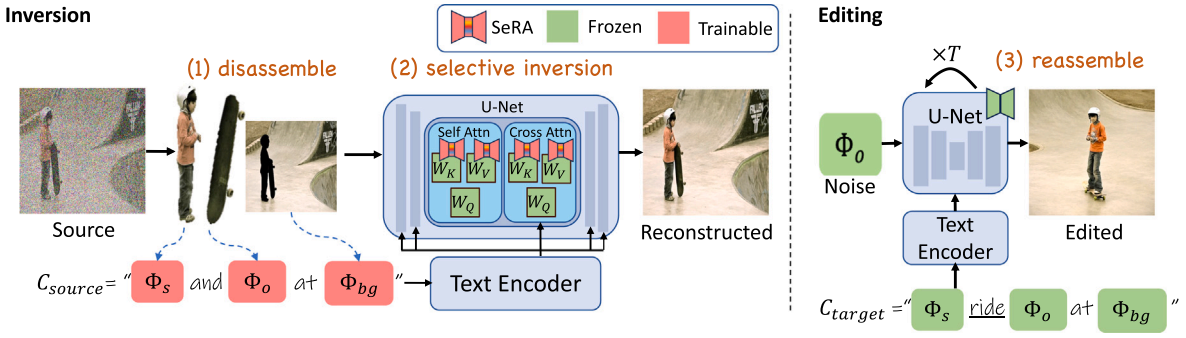


Fig. 2. Overview of the INTERACTEDIT framework. HOI components are disassembled into subject, object, and background descriptors during inversion (Section 3.3). SeRA regularization enables non-rigid edits by capturing essential identity features while suppressing overfitting to structural details (Section 3.4). Selective inversion preserves pretrained interaction priors while adapting to the source image's identity (Section 3.5). Editing reassembles these components with the target interaction, using SeRA weights delta to guide the diffusion model (Section 3.6).

trade-off between modifying the HOI and maintaining subject and object identity, all within a coherently adapted visual context.

Achieving this balance is non-trivial, as people, objects, and scene context are tightly interdependent and must remain visually coherent after editing. Yet, most modern image editors [8] focus on manipulating attributes, objects, or styles under a fixed layout assumption. This causes them to be ill-suited for interaction-level editing. **Interaction editing, by contrast, requires substantial structural changes**, such as reposing bodies, relocating objects and sometimes modifying the background, while preserving subject and object identity. For example, transforming *a person blowing a cake* into *eating the cake* (see Fig. 1) involves altering the action, adjusting body posture, and repositioning the object, all without disrupting scene coherence or visual likeness. Moreover, existing methods are not designed to operate without reference images or paired supervision [9]. This gap highlights the need for a new task formulation that supports flexible, identity-preserving edits directly from a single image and a target prompt.

To this end, we define the *reference-free HOI editing* task, which addresses four key challenges:

a) **Intricate interaction relationships**: subjects, objects, and actions are tightly coupled spatially & semantically;
b) **Rigid image structure**: existing editors rigidly preserve the original structure, limiting their ability to deform poses and layout changes required for new interactions;
c) **Pretrained-prior corruption**: unconstrained inversion risks overwriting pretrained interaction priors, thereby reducing generalizability to new target interactions;
d) **Missing standard dataset and metrics**: lack of a standard benchmark for a fair and reproducible comparisons.

To tackle these challenges, we propose INTERACTEDIT (Fig. 2), a novel framework that requires only a single source image and a target prompt. First, the image is decomposed into subject, object, and background tokens, facilitating the disentanglement of intricate HOI relationships. To support reference-free editing, it is essential to retain the model's pretrained understanding of interactions. We introduce *Selective Inversion*, which preserves the pretrained Query weights W^Q and adapts only the Key and Value weights W^K, W^V, thereby maintaining the original attention "question space" while allowing controllable interaction modifications. Building on this, we propose *Selective-Rank Adaptation* (SeRA) to guide each attention weight update within a structured, low-rank subspace, with each direction modulated by a learnable gating scalar. The goal of this design is to retain the pretrained action vocabulary while enabling identity-aware adaptations, striking a balance between interaction transformation and subject and object identity consistency.

To support a fair and reproducible evaluation, we also introduce IEBENCH, a dedicated benchmark for HOI editing. It systematically assesses HOI editability across a diverse set of source–target interaction

pairs and reports performance using the Editability–Identity score (EI), a new metric that jointly quantifies editing success and identity preservation.

As a summary, our contributions are:

(i) We formally define the reference-free HOI editing task and introduce IEBENCH, a comprehensive HOI benchmark, featuring diverse evaluation scenarios, as well as a new metric (i.e., Editability–Identity score (EI)) that jointly assesses HOI editability and identity preservation.
(ii) We present INTERACTEDIT, a novel framework for editing HOI from a single image. Specifically, it disassembles HOI components and combines *Selective Inversion* with SeRA to preserve pretrained action priors while enabling identity-consistent interaction edits.
(iii) Experiments show state-of-the-art performance where INTERACTEDIT achieves the best balance between HOI editability and identity consistency, establishing a strong foundation for future research in HOI editing.

A preliminary version of this work was previously disseminated as an arXiv preprint titled "InteractEdit: Zero-Shot Editing of Human-Object Interactions in Images" [10] prior to this submission.

2. Related work

2.1. Human-object interactions

Human-Object Interactions (HOIs) are typically represented as triplets ⟨human, action, object⟩ with corresponding bounding boxes for the human and the object. HOI detection aims to localize these entities and classify their interactions. Recent works [5,11–17] have improved generalization, particularly for rare and unseen interactions. HOI generation [6,7,18–22] is a recent field focused on generating images from given HOI triplets. InteractDiffusion [6] employs diffusion models to generate images matching the specified interactions and bounding boxes. While valuable for data augmentation and creative applications, HOI generation does not address editing interactions in existing images, where the goal is to modify the interaction while preserving original identity. Our proposed framework fills this gap by enabling interaction editing in images. It modifies existing HOIs while preserving the visual integrity of the original subject and object.

2.2. Diffusion models

Diffusion Models [23,24] generate high-quality images by iteratively denoising random noise and have proven to be more stable than GANs [25]. Conditional diffusion models further guide generation using text prompts [26–28], layout [29–33], personalized embeddings [34] and visual conditions [35,36]. Text-conditioned diffusion is itself a form of multimodal learning, where heterogeneous signals must be aligned and fused rather than merely concatenated [37,38]. While recent advances enable HOI-aware generation [6], editing existing human-object interactions while maintaining identity coherence remains largely unexplored. To bridge this gap, we leverage text-conditioned diffusion models, which are uniquely suited for flexible, controllable image manipulation.

2.3. Image editing diffusion models

Image Editing Diffusion Models [39,40] span single- and multi-subject customization, local editing, text-guided editing, and image translation [8]. Subject customization [41–50] generates or modifies images with specific subjects using a small set of reference images. Local editing [51–54] includes object changes [55] (adding, removing, replacing objects), attribute modifications (colour or texture), spatial transformations [53,56] (translation, scaling, warping), and inpainting [57]. Text-guided image editing [58–65] leverages natural language prompts to direct image modification, enabling a range of edits. Image translation [66,67] involves transforming images across domains (e.g., from sketches to photorealistic images) or applying style transfers. Mainstream customization methods typically require multiple images per subject, limiting their practical use. Break-A-Scene [68] showed that multiple visual concepts can be learned from a single image, which is crucial for real-world image editing where multiple reference images per concept are rarely available.

Although some of the above methods share building blocks with InteractEdit (e.g., concept decomposition, textual inversion, LoRA adaptation), they exhibit failure modes specific to HOI editing and serve different purposes from ours. DreamBooth-style personalization [69] adapts model weights but tends to overfit to the source image's structural information (original pose and spatial arrangement), making it unable to adapt to a new target interaction that requires large non-rigid changes. Textual Inversion [70] is too capacity-limited, as a few learned tokens cannot faithfully capture the appearance identity of both subject and object simultaneously. Break-A-Scene [68] decomposes concepts for multi-concept generation from scratch, rather than for preserving identities across a non-rigid interaction change. In contrast, InteractEdit explicitly separates these concerns: a subject-object-background decomposition disentangles the deeply coupled identities in an HOI image, while adapting model weights provides the representational capacity of DreamBooth without its structural overfitting. Furthermore, unlike prior parameter-freezing strategies (e.g., ControlNet [35], IP-Adapter [34]) that freeze parameters as a lightweight conditioning shortcut, our Selective Inversion is an interaction-aware design that follows directly from the algebraic structure of cross-attention (see Section 3.5): freezing W_Q preserves the pretrained interaction-aware spatial routing, while adapting W_K, W_V binds the new identity tokens to the prompt.

2.4. Interaction editing

Interaction editing is fundamentally distinct. While conventional editing modifies object attributes or positions within a largely static structure, interaction editing aims to change only the interaction while preserving subject and object identity. Person in Place [55] places humans onto a given background image rather than editing an existing interaction, making it unsuitable for direct comparison to HOI editing tasks. Recent advanced instructional image editing methods [71,72] excel at attribute or style manipulation but struggle with the flexible, non-rigid changes required by HOI edits. The concurrent HOIEdit [9] tackles HOI editing using multiple reference images of target interactions, making it less practical where such references are unavailable.

In contrast, our method enables *reference-free* HOI editing by leveraging the model's pretrained interaction knowledge, removing the need for reference images of target interactions. This makes our approach both practical and efficient. To the best of our knowledge, this is the ***first work*** to systematically address reference-free interaction editing.

2.5. Low-rank adaptation

LoRA [73] has been widely used in image editing for efficient fine-tuning [65], concept customization [74], and style transfer [75]. We extend LoRA with Selective-Rank Adaptation (SeRA), introducing per-rank gating to steer the inversion process more effectively. Unlike LoRA, our SeRA modulates each rank direction to capture essential identity attributes for reconstruction while suppressing overfitting to structural details. This adaptive mechanism is essential for flexible non-rigid interaction edits, a core requirement for effective reference-free HOI editing as pursued in this work.

3. Method

Given an input image I_{source} depicting an initial HOI as a triplet $\langle s, i_{\text{original}}, o \rangle$, where s is subject, o is object and i_{original} is their original interaction, our goal is to semantically modify this interaction to a new, specified interaction i_{target}. This requires balancing the preservation of subject and object identity with sufficient structural adaptability in editing the interaction and adapting the background as necessary.

The resulting image I_{target} should depict the new HOI $\langle s, i_{\text{target}}, o \rangle$ while maintaining realism and coherence.

3.1. Preliminary

Image editing generally involves two stages: *inversion* and *editing*. The inversion algorithm F_{inv} encodes an input image into a set of **descriptors** $\mathbf{\Phi}$ that capture the key identity and structural information needed for reconstruction. These descriptors are then used in the editing stage to generate the desired content. Formally, the inversion is defined as:

$$\mathbf{\Phi} = F_{\text{inv}}(I_{\text{source}}, C_{\text{source}}), \tag{1}$$

where C_{source} represents the original text prompt. During the editing stage, the edit function F_{edit} integrates $\mathbf{\Phi}$ into the base model to produce an image that retains visual identity while adhering to the new instruction:

$$I_{\text{target}} = F_{\text{edit}}(\mathbf{\Phi}, C_{\text{target}}), \tag{2}$$

where C_{target} is the target prompt.

Break-A-Scene [68] is an inversion technique that bridges Textual Inversion [70] and DreamBooth [69] for single-image multi-concept extraction. Technically, given an input image I_{source} and a set of N masks $\{M_n\}_{n=1}^{N}$ representing distinct concepts, it extracts the token descriptors $\{\Phi_n\}_{n=1}^{N}$, which can be incorporated into the target text prompts to generate new instances of these concepts. Additionally, it learns weight deltas ΔW to retain the source image information and improve reconstruction. While effective at disentangling static visual concepts, Break-A-Scene [68] struggles with non-rigid edits like HOI editing, which require modifying the spatial relationships between subjects and objects, as well as altering the subject's pose and the object's placement. Our approach, described next, enhances the Break-A-Scene with a regularized inversion process designed to better support non-rigid HOI edits.

3.2. InteractEdit: overall framework

We innovatively design InteractEdit (see Fig. 2) as a *reference-free* HOI-editing framework that leverages pretrained interaction priors to guide target edits without relying on external references. The core idea is to retain the model's existing HOI understanding while enabling flexible transformation. In other words, InteractEdit performs an inversion that learns weight deltas ΔW to encode the source image's identity while preserving the model's pretrained interaction priors. Specifically, we introduce *Selective Inversion* and Selective-Rank Adaptation (SeRA), which are designed to constrain updates to identity subspaces. This helps to retain the subject and the object identity while still allowing the targeted HOI transformation. The next section presents the decomposition of HOI to disentangle intricate HOI relationships, followed by a description of how *Selective Inversion* and SeRA are designed to leverage interaction priors and retain subject and object identity. Finally, we outline the overall HOI editing process.

3.3. Disassemble human-object interaction

Editing HOI in images is challenging due to the tightly coupled relationships between the human, object, and background. Directly editing these interdependent components often leads to one of two critical issues: either the intended interaction fails to transform, or the visual identity is lost. To address this issue, we **decompose** each scene into three elements: human (subject), object, and background, treating them as three distinct concepts, and represent them as a set of learnable token descriptors Φ_s, Φ_o, and Φ_{bg}, following Break-A-Scene [68]. These embeddings are incorporated into text prompts and optimized during inversion. Each concept is associated with a mask M_n, where $n \in s, o, bg$, which defines its spatial region within the image. At each optimization step, a subset of $k \leq N$ concepts is randomly selected, and a corresponding text prompt C_{source} is constructed. For example, if the subject and background are chosen, the text prompt takes the form: "a photo of $[\Phi_s]$ and $[\Phi_{\text{bg}}]$."

A masked version of standard diffusion loss [68] is then applied, ensuring that only the pixels covered by the selected concept masks contribute to optimization:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{\mathbf{z},\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon \odot M_U - \epsilon_\theta(\mathbf{z}_t, t, C_{\text{source}}) \odot M_U\|_2^2\right] \tag{3}$$

where $\mathbf{z}_t$ is the noisy latent at timestep t, C_{source} is the text prompt, M_U is the union of the corresponding masks, ϵ is the added noise, and, $\epsilon_\theta(\cdot)$ is the denoising network.

To further disentangle different concepts, an additional loss is imposed on the cross-attention maps to explicitly align each concept token Φ_n with its corresponding spatial region defined by the mask M_n, where $n \in s, o, bg$ (subject, object, and background). The cross-attention loss is defined as:

$$\mathcal{L}_{\text{attn}} = \mathbb{E}_{\mathbf{z},t,n}\left[\|\text{CA}_\theta(\Phi_n, \mathbf{z}_t) - M_n\|_2^2\right], \tag{4}$$

where $\text{CA}_\theta(\Phi_n, \mathbf{z}_t)$ is the cross-attention map between the token Φ_n and the noisy latent $\mathbf{z}_t$. Cross-attention maps are extracted at a 16×16 resolution, averaged across all corresponding U-Net layers, and normalized to the range (0,1). This regularization ensures that each learned concept token effectively corresponds to its spatial region, preventing entanglement between different concepts.

However, unconstrained updates as in [68], cause these descriptors to overfit to the structural details of the source image, such as spatial relationships, arrangement, and pose. In other words, this leads to the memorisation of specific configurations rather than learning generalizable representations. More critically, it corrupts the model's interaction priors, undermining its adaptability to new interactions. As a result, the generated outputs often fail to reflect the intended modifications, posing a significant barrier to effective HOI editing (see Table 4 and Fig. 7 for results).

3.4. Regularizing inversion via SeRA

To mitigate this, we introduce SeRA, a specialized LoRA [73] variant that integrates a learnable gating vector G between the low-rank matrices A and B (see Fig. 3). Technically, the gate $G \in \mathbb{R}^{r\times r}$ modulates each update direction, allowing the model to learn how many and which rank components to activate during inversion. Combined with the disentangled representation (from Section 3.3), this prevents the LoRA from overusing its capacity to memorise source-image structure, while retaining enough adaptation for HOI transformations such as pose and subject–object spatial relations.

Formally, a SeRA-adapted weight matrix W is defined as:

$$W = W_{\text{original}} + \Delta W, \quad \Delta W = AGB^\top, \tag{5}$$

where $W \in \mathbb{R}^{d_1\times d_2}$, $A \in \mathbb{R}^{d_1\times r}$, $B \in \mathbb{R}^{d_2\times r}$, r is the rank hyperparameter, and $G \in \mathbb{R}^{r\times r}$ is a diagonal, learnable matrix that gates the contribution of each rank direction. In practice, we apply a sigmoid to each diagonal

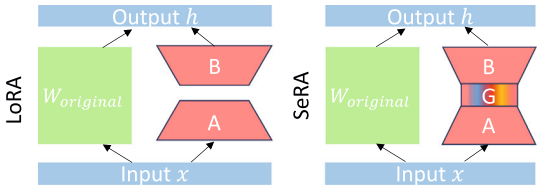


Fig. 3. Selective-Rank Adaptation (SeRA): It extends LoRA by integrating a learnable gating vector G between low-rank matrices A and B, enabling selective activation of rank directions during weight updates. This constrains adaptation capacity to reduce overfitting to source-image structure while supporting adaptive interaction edits.

entry of G, i.e., $G \leftarrow \sigma(G)$, so that every gate lives in $(0, 1)$ and softly controls each rank direction. We apply SeRA in the U-Net backbone, as detailed in the following section.

To ensure each gate in G commits to either opening or closing its corresponding subspace, rather than hovering at an indecisive midpoint, we introduce an uncertainty loss that actively repels gates from 0.5. Concretely, for a diagonal gating matrix $G \in \mathbb{R}^{r\times r}$ with entries G_{jj}, we define:

$$\mathcal{L}_{\text{SeRA}} = \frac{1}{r}\sum_{j=1}^{r} G_{jj}\left(1 - G_{jj}\right). \tag{6}$$

Since the term $G_{jj}(1 - G_{jj})$ attains its maximum at $G_{jj} = 0.5$ and falls to zero at the endpoints 0 and 1, minimizing $\mathcal{L}_{\text{SeRA}}$ drives each gate away from uncertainty and toward a clear "on" or "off" state.

3.5. Selective inversion for prior retention

The U-Net used in diffusion models comprises multiple blocks arranged in an encoder–decoder structure. Each block contains self-attention and cross-attention mechanisms to capture relationships within the image features and between the image and conditioning inputs, respectively. In each attention layer, input features are transformed into Query (Q), Key (K) and Value (V) representations. The attention mechanism is described as

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V,$$

where $Q = XW^Q$, $K = XW^K$, and $V = XW^V$ are linear transformations of the input X using weight matrices W^Q, W^K, and W^V, respectively, and d_k is the dimensionality of the Key vectors used for scaling. Q encodes what features a token is "looking for," K represents what features a token offers to be attended to, and V carries the actual information used to generate the output.

In HOI editing, our goal is to leverage the pretrained model's interaction knowledge while learning to encode and maintain the visual traits of the source image. Rather than assigning trainable and frozen projections heuristically, we motivate the choice from the algebraic structure of attention. The attention operation can be factorised into a *structural* factor and a *content* factor:

$$\underbrace{A = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)}_{\text{structure: } where \text{ attention is routed}}, \qquad \underbrace{O = AV}_{\text{content: } what \text{ is placed there}}. \tag{7}$$

Two observations follow immediately from Eq. (7). First, V does *not* enter the routing matrix A at all; it only modulates the *content* that is read out at each spatial location. Second, A is determined jointly by Q and K: Q comes from the image-side latent X_{img} and acts as the spatial query, while K comes from the text embedding X_{txt} and provides the per-token binding key for routing. In cross-attention, the asymmetry is sharper still: Q is image-side and K, V are text-side, so the newly introduced concept tokens for the disassembled subject, object, and background can only enter the attention through W^K and W^V.

This factorisation motivates a design for which projections to adapt and which to freeze during inversion:

- **Train W^V (content lever).** Since V is the only factor in AV that is not shared with A, adapting W^V provides a direct content pathway for writing the inverted subject/object/background appearance into the output without perturbing the routing.
- **Train W^K (binding key).** Adapting W^K helps make the newly inverted concept tokens addressable by the image queries; otherwise their attention column $A[:, \text{new}] \propto \exp(Qk^{\top})$ with $k = X_{\text{txt}}W^K$ remains generic, and the new content in V is not routed to the intended spatial region.
- **Freeze W^Q (pretrained spatial routing).** W^Q produces the image-side queries that select *where* text tokens attend, and the pretrained W^Q already encodes an interaction-aware spatial routing prior. Freezing W^Q preserves this pathway rather than fitting it to the pose and spatial layout of the single source image, which we expect to mitigate the "structural locking" that otherwise hinders re-routing toward the target interaction (Fig. 8).

Our hypothesis is therefore that freezing W^Q while adapting W^K, W^V should preserve the general spatial routing while still allowing the new identities to bind to the prompt. Leaving both W^Q and W^K trainable instead removes the constraint on A, so the adaptation is free to fit, and potentially *memorise the source attention map*; we confirm empirically in Section 5.4.1 that this fully-trainable configuration performs worst. For the frozen-W^Q case, Lemma 1 further shows that the row-space of the attention-score matrix $S(X)$ remains contained in the row-space induced by the pretrained query map, for any update of W^K.

We apply the same Q-frozen/K, V-trainable schedule consistently to both self- and cross-attention: in self-attention the same decomposition holds with Q, K, V all derived from the image latent, and freezing W^Q analogously preserves the pretrained intra-image spatial relations. These updates are parameterized with low-rank SeRA modules. Specifically, the pretrained weights (W_{original}) remain frozen, while only the SeRA parameters (A, B and G) are optimized. Collectively, the SeRA-adapted weights across all U-Net blocks constitute Φ_{SeRA}, capturing image-specific adjustments that faithfully reconstruct the source while retaining pretrained interaction priors.

The overall loss combines the above masked reconstruction loss (Eq. 3), the cross-attention alignment loss (Eq. 4) and the SeRA loss (Eq. 6):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{attn}}\mathcal{L}_{\text{attn}} + \lambda_{\text{SeRA}}\mathcal{L}_{\text{SeRA}}, \tag{8}$$

where $\lambda_{\text{attn}} = 0.01$ following Break-A-Scene [68] and $\lambda_{\text{SeRA}} = 0.01$.

3.6. HOI editing

To implement selective inversion, we extract descriptors as:

$$\mathbf{\Phi} = [\Phi_s, \Phi_o, \Phi_{\text{bg}}, \Phi_{\text{SeRA}}] = F_{\text{inv}}(I_{\text{source}}, C_{\text{source}}), \tag{9}$$

where Φ_s, Φ_o, Φ_{bg} are the descriptors representing the subject, object, and background, respectively, and Φ_{SeRA} denotes the SeRA weight deltas learned during inversion.

During the editing phase, we generate an image that depicts the target interaction while preserving the original identity. To do this, we reassemble the descriptor tokens (Φ_s, Φ_o, and Φ_{bg}) into a new target prompt, replacing the original action i_{original} with the target interaction i_{target}:

$$C_{\text{target}} = \text{"a photo of } \left[\Phi_s\right]\ i_{\text{target}}\ \left[\Phi_o\right]\ \text{at}\ \left[\Phi_{\text{bg}}\right]\text{."}$$

Using this prompt, the diffusion model ϵ_θ, along with selectively inverted weights Φ_{SeRA}, generates the edited image I_{target} via an iterative denoising process starting from random noise $\mathbf{z}_T$. At each timestep $t = T, T-1, \ldots, 1$, the model predicts the denoised image:

$$\mathbf{z}_{t-1} = \epsilon_\theta(\mathbf{z}_t, t, C_{\text{target}};\ \Phi_{\text{SeRA}}). \tag{10}$$

The final image I_{target} is obtained as $\mathbf{z}_0$. This process realizes Eq. (2), where $\mathbf{\Phi}$ includes the descriptor tokens (Φ_s, Φ_o, Φ_{bg}) and SeRA weights Φ_{SeRA}.

In this process, the selectively inverted SeRA weights Φ_{SeRA} and the descriptor tokens (Φ_s, Φ_o, Φ_{bg}) work together to balance preserving the visual identity from the source image and enabling sufficient flexibility for interaction edits, while the pretrained model parameters provide interaction priors. This synergy ensures the generated image reflects the new interaction i_{target} as specified by C_{target}, without significantly compromising the original identity.

Table 1
Objects and their candidate target interactions in IEBENCH and IEBENCH-L. The original IEBENCH objects are listed in the top block; IEBENCH-L extends this set with the additional objects and actions.

Subset	Object	Target Interactions
IEBENCH	Broccoli	cut, eat, hold, smell, wash
	Dining table	clean, eat at, sit at
	Skateboard	hold, jump, ride, sit on
	Chair	hold, lie on, sit on, stand on
	Book	carry, hold, read
	Snowboard	hold, jump, ride
	Surfboard	hold, jump, ride, sit on
	Sports ball	catch, dribble, hit, hold, kick, throw
	Cake	blow, cut, eat, hold, make
	Horse	feed, kiss, ride, walk, wash
	Dog	feed, groom, hug, walk, wash
	Pizza	cut, eat, hold, make, pick up
	Cat	feed, hold, hug, kiss, wash
IEBENCH-L	Bench	sit on, lie on
	Cup	hold, drink with, sip, fill, smell
	Banana	hold, eat, cut
	Bottle	drink with, open, hold
	Couch	sit on, lie on
	Teddy bear	hold, hug, carry
	Donut	hold, eat, pick up
	Sandwich	hold, eat

4. Interaction editing benchmark (IEBench)

4.1. Benchmark

A rigorous benchmark is essential for properly evaluating HOI editing task. While HOIEdit [9] made initial progress, it has two key limitations. First, its evaluation is limited to a small-scale benchmark covering only 3 actions. Second, it relies on text-image and image–image similarity metrics, which are unreliable for assessing HOI edit correctness, as CLIP is known to underperform in the action recognition task [76].

To address these shortcomings, we introduce IEBENCH,[1] a new benchmark specifically designed to evaluate both intended interaction transformation and identity preservation in the HOI editing task. IEBENCH is constructed from the test subset of HICO-DET dataset [77] by selecting source images and target interactions that: (i) the source image must depict a single HOI instance, (ii) the target interaction must be plausible for the object, and (iii) the target interaction must produce significant visual changes. We define criterion (iii) operationally: a ⟨source image, target interaction⟩ pair is retained only if realizing the target requires the person to physically reconfigure their limbs or body relative to the object. For example, we retain *hold* → *eat* broccoli (the arm raises the object to the mouth) and *hold* → *ride* a skateboard (the board moves from the side to underfoot), while excluding near-synonymous verb swaps that leave the same limbs in the same contact with the object, as well as attribute-only edits such as object colour or texture, since a method could satisfy these without changing the interaction at all.

The resulting benchmark comprises 28 source images covering 25 actions and 13 objects, forming a total of 100 unique ⟨ source image, target interaction ⟩ pairs. The full set of actions includes: *feed, make, pick up, sit on, hit, ride, walk, cut, eat at, jump, throw, dribble, smell, kick, hug, eat, hold, catch, sit at, wash, stand on, kiss, groom, carry, and lie on*. Table 1 lists the complete set of objects and their paired target interactions. Each object is matched with multiple interactions to ensure diverse and meaningful interaction editing scenarios.

4.1.1. Extended benchmark (IEBENCH-L)

To evaluate generalization at a larger scale and broader interaction coverage, we additionally introduce IEBENCH-L, an extended and *more challenging* version of IEBENCH constructed under the same selection criteria. IEBENCH-L comprises **100 source images** (**3.6**× larger), spanning **30 unique actions** (**1.2**×) and **21 unique objects** (**1.6**×), and forming **326** unique ⟨source image, target interaction⟩ pairs (**3.3**×). Beyond scale, IEBENCH-L is intentionally harder: it covers a much wider distribution of subjects, objects, and viewpoints, and incorporates rarer verb–object compositions and more demanding non-rigid pose changes that stress-test both the editability and identity-preservation aspects of HOI editing. The newly added objects and their candidate target interactions are listed in the bottom block of Table 1. IEBENCH remains a strict subset of IEBENCH-L, ensuring backward compatibility with all results reported on IEBENCH. We re-evaluate INTERACTEDIT together with the top-performing baselines on IEBENCH-L to verify that our conclusions hold under this larger and more demanding setting; the full results are reported in Section 5.3.

4.2. New metric

Traditional metrics like CLIP-T, which measure text–image alignment between a prompt and the generated image, are inadequate for HOI editing. CLIP [78] is known to struggle with verb semantics [79,80]. As shown in Fig. 4, very different human–horse interactions (e.g., stand, wash, ride) yield nearly identical similarity scores with the prompt "person rides horse," demonstrating CLIP-T's invariance to verbs. To analyze this, we further test CLIP-T's sensitivity to action (verb) versus object mismatches. Verb mismatches (blue) yield extremely narrow margins clustered near zero, while noun mismatches (orange) produce broader margins. This confirms CLIP is primarily noun-sensitive but weak at distinguishing verbs, highlighting why CLIP-T cannot serve as a reliable HOI metric and motivating the need for our HOI-specific alternatives. Further details are given in Appendix B.

To complement this benchmark, we introduce new evaluation metrics specifically designed for HOI editing task. These metrics quantify the trade-off between intended interaction transformation correctness and identity preservation.

(i) **HOI Editability, HE** quantifies editing success by determining whether the target interaction is present in the edited image. Leveraging PViC [11], a state-of-the-art HOI detector, each generated image is assigned a score of one if the target interaction is detected, and zero otherwise. The final HE score is computed as the mean detection rate over all edited samples.

(ii) **Identity Consistency** assesses how well the subject and object identities are preserved after editing. It is deliberately region-restricted, comparing only the segmented subject and object regions, so that the large pose and layout changes required by an interaction edit are not themselves penalized.

Let I be a source image and $\hat{I}$ its edited counterpart. For a concept $c \in \{s, o\}$ (subject, object), we define the region-extraction operator

$$\mathcal{R}_c(I) = I \odot \mathrm{SAM}\big(I, \mathrm{GDINO}(I, \tau_c)\big), \tag{11}$$

which grounds the concept using GroundingDINO [81] with the text query τ_c (the subject or object noun of the HOI triplet), segments it with SAM [82], and returns the masked region. Let $f(\cdot) \in \mathbb{R}^d$ denote the DINOv2 [83] embedding of a masked region, chosen for its strong object-centric and fine-grained representation. The per-concept identity similarity is then the cosine similarity between the source and the edited region:

$$\mathrm{sim}_c\big(I, \hat{I}\big) = \frac{\big\langle f(\mathcal{R}_c(I)),\, f(\mathcal{R}_c(\hat{I}))\big\rangle}{\big\|f(\mathcal{R}_c(I))\big\|_2 \cdot \big\|f(\mathcal{R}_c(\hat{I}))\big\|_2}. \tag{12}$$

Each edited image is scored by the average of sim_s and sim_o, and the final Identity Consistency score is computed as the mean over

[1] We will release this benchmark publicly.



Fig. 4. CLIP-T score fails to distinguish verb-specific interaction differences.

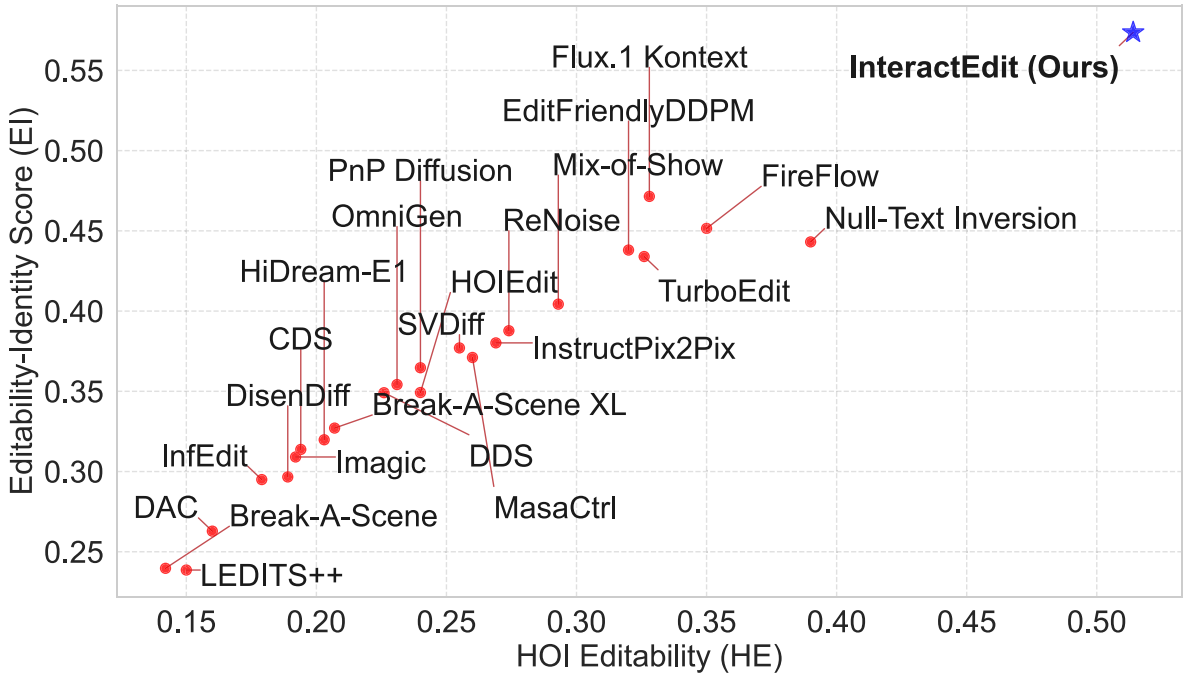


Fig. 5. Our method achieves the best results for both HOI editability and Editability–Identity score.

all edited samples. If GroundingDINO fails to ground concept c in the edited image, we set $\text{sim}_c = 0$ rather than excluding it from the average, so that a method cannot obtain a high identity score by removing the subject or the object altogether.

(iii) **Editability–Identity Score,** EI quantifies the trade-off between HE score and Identity Consistency via the harmonic mean, analogous to the F_1 score [84]. This formulation ensures a balanced evaluation by penalizing low performance in either dimension:

$$\text{EI} = \frac{2 \times \text{HOI Editability} \times \text{Identity Consistency}}{\text{HOI Editability} + \text{Identity Consistency}}. \tag{13}$$

(iv) **Human Preference Score v2 (HPSv2)** [85] is a perceptual evaluation model [86–88] trained on extensive human rankings to quantify the alignment of generated images with human aesthetic and quality preferences. We leverage this metric to assess the fine-grained visual realism of the edited HOI scenes, which automated structural metrics may overlook. While HE and EI scores ensure semantic interaction correctness and identity consistency, **HPSv2** provides a necessary complementary perspective by evaluating the overall photorealism of the generated results.

5. Experiments

We conducted experiments at 512 × 512 resolution, basing our method on Stable Diffusion XL [28]. Following Break-A-Scene [68], we use a two-stage optimization approach. In the first stage, we optimize only the descriptor tokens $\Phi_s, \Phi_o, \Phi_{\text{bg}}$ for 1000 steps with a learning rate of 5e-4. In the second stage, we jointly optimize the SeRA weights Φ_{SeRA} and the descriptor tokens for 200 additional steps with a learning rate of 1e-4. Additional details are in Appendix A.

5.1. Qualitative results

Fig. 6 compares our edits with those of leading baselines, illustrating how each method trades off interaction change versus identity preservation. NTI and InstructPix2Pix manage to alter actions but severely distort the person or object. InfEdit, Flux.1 Kontext, and Break-A-Scene instead lock in identity perfectly yet leave poses and layouts essentially unchanged. For example, when asked to *"wash the horse,"* Flux.1 Kontext still depicts the rider in the original "ride horse" pose, resulting in almost no interaction change.

Our approach strikes the balance: SeRA's gated low-rank updates constrain inversion to identity-relevant subspaces, avoiding overfitting

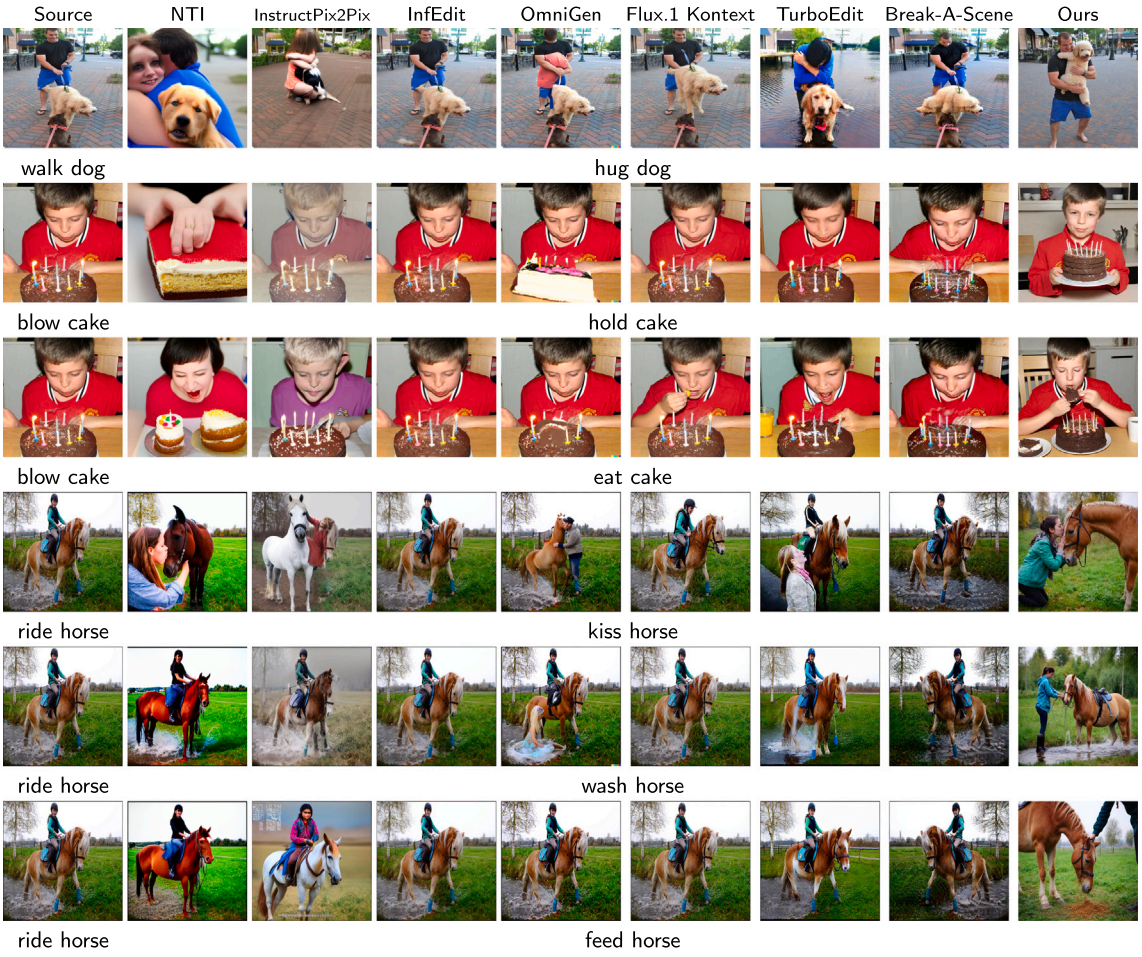


Fig. 6. Qualitative comparison with SOTAs. The source column shows the original image and its associated interaction. Our method demonstrates the most balanced results among all methods. Additional qualitative results are provided in Appendix D.

to the source structure. Selective inversion freezes the model's query layers so that pretrained interaction priors remain intact. This synergy unlocks substantial non-rigid edits like re-posing arms and repositioning objects, without sacrificing who or what is in the scene. The crisp "hold cake" and dynamic "kiss horse" examples best illustrate this effective trade-off (see Appendix D for more).

5.2. Quantitative results

For each pair of source image and target interaction in IEBench, we generate 10 outputs using different random seeds, totalling 1000 edited images. Table 2 and Fig. 5 compare our method against 23 existing baselines in terms of HOI editability and Editability–Identity scores. InteractEdit outperforms all others, advancing the state of HOI editing.

Our InteractEdit achieves an HOI Editability of **0.514**, a **31.8%** relative gain over the previous best (NTI). On the composite EI score, which harmonically combines editability and identity consistency, we reach **0.573**, outperforming the next-best method (Flux.1 Kontext) by **21.7%**. Furthermore, our framework secures the highest HPSv2 score of **0.2640**, surpassing FireFlow (0.2530).

These results demonstrate that our approach not only improves interaction accuracy but also maintains a strong trade-off between posing new actions and preserving visual identity, while also advancing perceptual quality. We attribute these gains to our SeRA-based inversion regularization, which prevents overfitting to the source image, and our *Selective Inversion* strategy, which preserves target interaction priors from the pretrained model.

5.3. Quantitative results on IEBench-L

To verify the generalizability and the scalability beyond IEBench, we re-evaluate InteractEdit together with the top-performing baselines on the extended IEBench-L benchmark (Section 4.1.1). For each ⟨source image, target interaction⟩ pair, we generate **5** samples with different random seeds, totalling **1630** edited images per method. Table 3 reports HOI Editability (HE), Editability–Identity (EI), Identity Consistency, and HPSv2 on IEBench-L.

IEBench-L is intentionally harder than IEBench (broader object distribution, rarer HOIs, and more demanding non-rigid pose changes), providing a stricter test of HOI editing capability. Despite this, InteractEdit remains the top performer on the EI with **0.332**, **6.4%** relative gain over the next-best InfEdit (0.312); the HE with **0.498**, **38.7%** relative gain over the runner-up Null-Text Inversion (0.359); and the highest HPSv2 of **0.2625** (vs. FireFlow's 0.2549). On Identity Consistency alone, InfEdit attains the highest score (0.831); however, this comes at the cost of editability: InfEdit's HE of only 0.192, shows that it preserves identity largely by leaving the source pose unchanged, failing to actually realize the target interaction. This is precisely the trade-off EI is designed to capture, where the InteractEdit strikes the most effective balance between editability and identity preservation. This empirical result shows that InteractEdit transfers cleanly

Table 2
Comparison of INTERACTEDIT with existing baselines on the Editability–Identity score EI, HOI Editability and Identity Consistency using IEBENCH. We use HPSv2 [85] as a human-preference aligned perceptual metric to evaluate fine-grained realism. The **best** and second best are indicated.

Method	EI	HOI Editability	Identity Consistency	HPSv2
Null-Text Inversion [89]	0.443	0.390	0.513	0.2483
PnP Diffusion [90]	0.365	0.240	0.759	0.2520
MasaCtrl [91]	0.371	0.260	0.648	0.2212
HOIEdit [9]	0.349	0.240	0.641	0.2129
Imagic [60]	0.309	0.192	0.792	0.2417
CDS [92]	0.314	0.194	0.820	0.2303
DAC [65]	0.263	0.160	0.736	0.2135
LEDITS++ [64]	0.239	0.150	0.583	0.2043
InstructPix2Pix [58]	0.380	0.269	0.648	0.2178
ReNoise [93]	0.388	0.274	0.663	0.2312
TurboEdit [94]	0.434	0.326	0.649	0.2437
SVDiff [44]	0.377	0.255	0.723	0.2351
EditFriendlyDDPM [95]	0.438	0.320	0.694	0.2470
DDS [96]	0.349	0.226	0.767	0.2270
InfEdit [97]	0.295	0.179	**0.838**	0.2296
Break-A-Scene [68]	0.240	0.142	0.768	0.2296
Break-A-Scene XL	0.327	0.207	0.779	0.2367
Mix-of-Show [74]	0.404	0.293	0.652	0.2505
DisenDiff [98]	0.297	0.189	0.689	0.2234
OmniGen [99]	0.354	0.231	0.759	0.2120
HiDream-E1 [100]	0.320	0.203	0.753	0.2297
Flux.1 Kontext [72]	0.471	0.328	**0.838**	0.2427
FireFlow [101]	0.451	0.350	0.636	0.2530
InteractEdit (Ours)	**0.573**	**0.514**	0.649	**0.2640**

Table 3
Comparison of INTERACTEDIT with existing baselines on the more challenging IEBENCH-L benchmark, reporting Editability–Identity score (EI), HOI Editability, Identity Consistency, and HPSv2 [85] (a human-preference aligned perceptual metric for fine-grained realism). IEBENCH-L scales IEBENCH to 100 source images, 33 actions, and 21 objects with rarer verb–object compositions and harder non-rigid pose changes, causing absolute scores to drop across all methods. The **best** and second best are indicated.

Method	EI	HOI Editability	Identity Consistency	HPSv2
Null-Text Inversion [89]	0.296	0.359	0.252	0.2437
MasaCtrl [91]	0.243	0.245	0.241	0.2205
TurboEdit [94]	0.280	0.299	0.264	0.2426
EditFriendlyDDPM [95]	0.255	0.258	0.252	0.2438
InfEdit [97]	0.312	0.192	**0.831**	0.2332
Mix-of-Show [74]	0.260	0.282	0.241	0.2524
Flux.1 Kontext [72]	0.272	0.276	0.269	0.2471
FireFlow [101]	0.290	0.337	0.255	0.2549
InteractEdit (Ours)	**0.332**	**0.498**	0.250	**0.2625**

from smaller-scale IEBENCH (Table 2) to the broader, more demanding 100-image, 30-action, 21-object setting of IEBENCH-L.

5.4. Ablation studies

HOI editing poses significant challenges, such as overfitting to the source image and ineffective use of pretrained interaction knowledge, limiting the adaptability to target interactions. To address these, we conduct an ablation study to assess the contributions of our *Selective Inversion* strategy and SeRA. The results are summarized in Table 4 and visualized in Fig. 7. The baseline refers to SDXL with DreamBooth [69] and Textual Inversion [70].

The baseline configuration greatly overfits the source image, yielding a very low EI (0.252) and HE (0.150) scores. Simply disassembling HOI components raises EI to 0.327 and HE to 0.207, showing that concept split alone is insufficient without suitable regularization. Introducing

Table 4
Ablation study of INTERACTEDIT. Dis., SI, and SeRA represent Disassemble HOI, Selective Inversion, and regularization with SeRA.

Method	Dis.	SI	SeRA	EI	HOI Editability
B0: Baseline				0.252	0.150
B1: +Dis	✓			0.327	0.207
L1: LoRA			LoRA	0.454	0.331
S1: SeRA			✓	0.498	0.388
S2: SeRA + SI		✓	✓	0.508	0.410
A: AdaLoRA + SI +Dis.	✓	✓	AdaLoRA	0.525	0.502
S3: Full (Ours)	✓	✓	✓	**0.573**	**0.514**

standard LoRA across all parameters boosts EI to 0.454 and HE to 0.331, demonstrating that capacity control helps unlock interaction changes but still risks overwriting pretrained interaction priors. Applying SeRA, which selectively gates parameter adaptation, raises EI to 0.498 and HE to 0.388, underscoring the value of constrained, direction-wise updates that preserve pretrained priors. Adding *Selective Inversion* to SeRA yields a slight uptick (EI = 0.508 and HE = 0.410), confirming that freezing interaction queries enhances editability by preserving pretrained interaction knowledge. Finally, integrating disassembly with *Selective Inversion* and SeRA achieves the best performance (EI = 0.573 and HE = 0.514). For completeness, we also compare against AdaLoRA [102], which adaptively prunes ranks based on importance scores. As shown in Table 4, AdaLoRA has good editability (HE = 0.502) but achieves a lower EI (0.525) than SeRA. This is because rank pruning discards entire update directions, which can erase useful identity information, leading to less stable identity preservation. In contrast, SeRA softly regulates each rank through continuous gating, striking a more effective balance between editability and identity consistency.

To empirically validate **Selective Inversion**, we visualize the cross-attention maps in Fig. 8 for the target interaction token *"ride"* across different inversion strategies. In the **Full Inversion** baseline, the attention map remains spatially anchored to the original *"hold"* interaction region. This "structural locking" occurs because updating all weights causes the model to overfit to the source image's layout, thereby corrupting the pretrained interaction priors and preventing the intended transformation. In contrast, our **Selective Inversion** freezes the Query weights (W_Q), which preserves the model's pretrained "question space" and its fundamental interaction reasoning capabilities. As shown in the visualization, this constraint enables the attention map to successfully shift its focus to the new semantic structure required for the *"ride"* interaction, facilitating the successful rendering of the intended edit. This empirical evidence confirms that W_Q is the primary component responsible for structural retrieval and interaction logic, while updating only the Key (W_K) and Value (W_V) weights is sufficient for capturing the visual identities of the subject and object from the source image.

Fig. 8 shows one map per method, pooled over the cross-attention layers. To confirm that the effect is not an artefact of this pooling, Fig. 9 resolves the same comparison by depth, at the downsample, mid, and upsample blocks. The pattern is consistent throughout: under Full Inversion the attention for the target action token stays organised around the source interaction layout at the downsample, mid, and upsample blocks alike, whereas under Selective Inversion it is organised around the target layout.

5.4.1. Which cross-attention projection should be frozen?

To verify the selective inversion derived from Eqs. (7) and Lemma 1 in Section 3.5, we re-train INTERACTEDIT under every partition of $\{W_Q, W_K, W_V\}$ into trainable and frozen projections, keeping concept disassembly and SeRA fixed. Table 5 reports EI, HOI Editability, and Identity and Background consistency on IEBENCH.

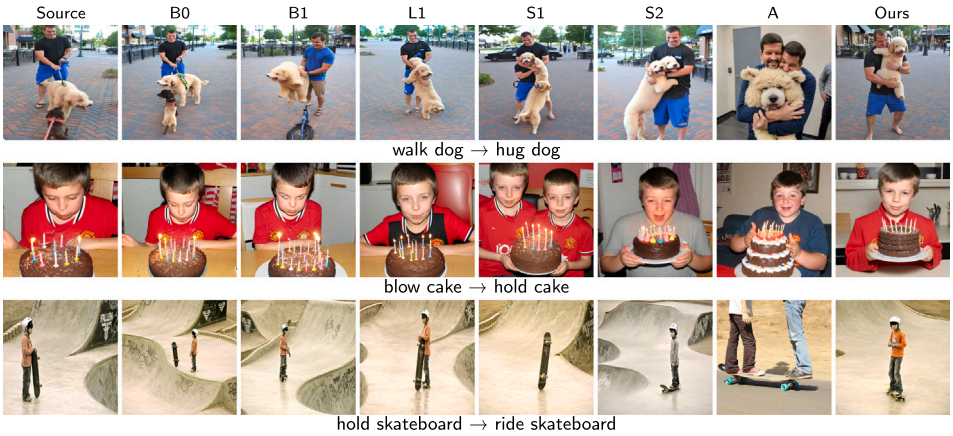


Fig. 7. Ablation study comparing the effects of various model components on HOI editing performance. Each row shows a source image and the edited outputs for a given interaction edit (e.g., walk dog → hug dog).

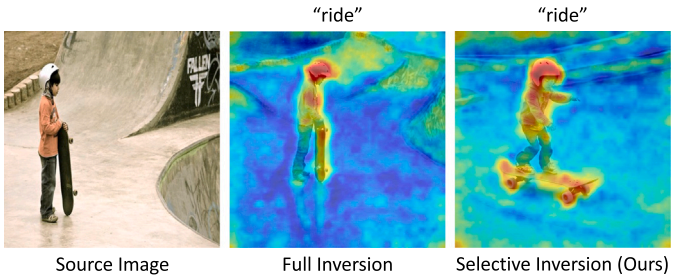


Fig. 8. Cross-attention maps on the action token "ride". In **Full Inversion**, the model overfits to the source structure, leaving the attention map stuck on the original interaction region (holding skateboard). In our **Selective Inversion**, freezing W_Q retains the pretrained "question space," enabling the attention to successfully shift to the new "riding" semantic layout. This validates the role of W_Q in preserving interaction reasoning capabilities during inversion.

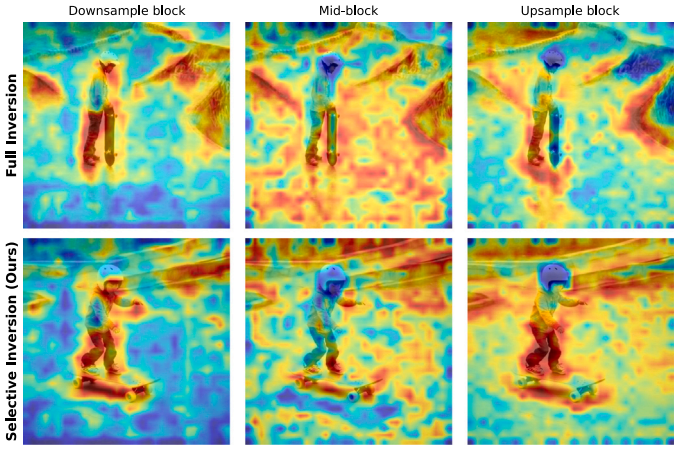


Fig. 9. Cross-attention on the action token *"ride"* at three U-Net depths. Under **Full Inversion** (top) the attention remains organised around the source *hold* configuration, at every depth, whereas under **Selective Inversion** (bottom) it follows the target *ride* configuration, covering the rider's body and the board underfoot. Warm colours denote higher attention; maps are averaged over heads and denoising timesteps.

Table 5
Ablation on which cross-attention projections to freeze during Selective Inversion. ✓ indicates the projection is *trainable*; × indicates it is *frozen*. Freezing only W_Q (our Selective Inversion) yields the best overall trade-off.

Configuration	W_Q	W_K	W_V	EI	HOI Editability	Identity	Background
None frozen (Full Inversion)	✓	✓	✓	0.496	0.449	0.555	0.634
Freeze W_K	✓	×	✓	0.558	0.502	0.628	0.596
Freeze W_V	✓	✓	×	0.528	0.489	0.574	0.509
Freeze W_K, W_V	✓	×	×	0.553	**0.540**	0.566	0.515
Freeze W_Q (Ours)	×	✓	✓	**0.573**	0.514	**0.649**	**0.641**

(i) Training W_Q and W_K jointly is strictly worse. The "None frozen" row drops to EI = 0.496 (7.7 points below ours): with both factors of A trainable, the optimizer memorises the source-image attention map, exactly as predicted, and exhibits the structural locking visible in Fig. 8. The high Background score (0.634) reflects pixel-level source reconstruction, not interaction editing. **(ii) Freezing W_V or W_K starves the inverted identity.** Freezing W_V removes the only content lever in $O = AV$ and collapses Background to 0.509—the worst of any configuration. Freezing W_K instead forces the new concept tokens to use the pretrained key projection, which has never seen these tokens, so image queries cannot selectively attend to them; EI settles at an intermediate 0.558. Both confirm that W_K and W_V are the necessary entry points for the inverted identity. **(iii) Freezing both W_K and W_V over-edits.** Training only W_Q attains the highest HOI Editability (0.540) by bending the routing toward generic pretrained interactions, but Identity (0.566) and Background (0.515) drop sharply: the inverted concept tokens have no path into the prompt. **(iv) Freezing only W_Q (ours) is best trade-off.** The configuration certified by Lemma 1 simultaneously attains the best EI (0.573), Identity (0.649) and Background (0.641), while remaining competitive on HOI Editability (0.514). No alternative partition dominates ours on more than a single metric.

The main effect of W_Q. Reading Table 5 to determine whether W_Q is trainable isolates its contribution independently of W_K and W_V. The four configurations that leave W_Q trainable achieve EI values of 0.496, 0.528, 0.553, and 0.558 (mean 0.534), all below the 0.573 obtained by freezing it. Marginalising over what happens to W_K and W_V, freezing W_Q is therefore worth +0.039 EI on average, and no W_Q-trainable partition matches the frozen-W_Q result even at its best.

Together with Lemma 1 and the attention visualization in Figs. 8 and 9, this ablation provides the missing empirical link: the subspace preserved by freezing W_Q is interaction-relevant, not merely a generic visual subspace.

Table 6
Sensitivity of InteractEdit to mask quality. InteractEdit was retrained under 10 perturbed mask conditions (bbox replacement and dilation/erosion/translation at ±5/15/30 px). The last row shows the grouped means. Performance remains within $\approx 1\sigma$ of the original baseline across all three metrics.

Mask condition	mask IoU	Editability-Identity (EI)	HOI Editability	Identity Consistency
Original (baseline)	1.00	0.560	0.579	0.542
Tight bounding box	0.50	0.573	0.671	0.500
Dilation + 5 px	0.83	0.570	0.618	0.528
Dilation + 15 px	0.64	0.511	0.513	0.509
Dilation + 30 px	0.49	0.538	0.605	0.484
Erosion −5 px	0.77	0.588	0.645	0.540
Erosion −15 px	0.56	0.517	0.526	0.509
Erosion −30 px	0.38	0.627	0.724	0.553
Translation 5 px	0.65	0.540	0.579	0.507
Translation 15 px	0.57	0.577	0.632	0.534
Translation 30 px	0.68	0.536	0.548	0.524
All perturbed (mean ± std)	0.62	0.553 ± 0.036	0.595 ± 0.070	0.520 ± 0.021

5.4.2. Sensitivity to mask quality

InteractEdit obtains its subject and object masks from an automatic SAM + GroundingDINO pipeline, raising the question of whether the method depends on the precision of this off-the-shelf segmenter. To investigate, we evaluate InteractEdit from scratch on 6 random IEBench scenes under 10 perturbed mask conditions covering two axes: (i) mask source—replacing the original mask with its tight bounding box, simulating the weakest practical prior a user could supply; and (ii) controlled geometric degradation of the SAM mask via dilation, erosion, or translation at ±5/15/30 px. Only the subject and object masks are perturbed; the background mask is recomputed as the complement to remain consistent. Across all 10 perturbed conditions, the perturbed mean stays within approximately one standard deviation of the original baseline on every metric: HOI Editability (0.579 → 0.595 ± 0.070), Identity Consistency (0.542 → 0.520 ± 0.021), and EI (0.560 → 0.553 ± 0.036) (Table 6). We conclude that InteractEdit is **insensitive to mask quality** and that the masks act as attention localizers, not pixel-level constraints, so a bounding box from an off-the-shelf detector is a sufficient prior.

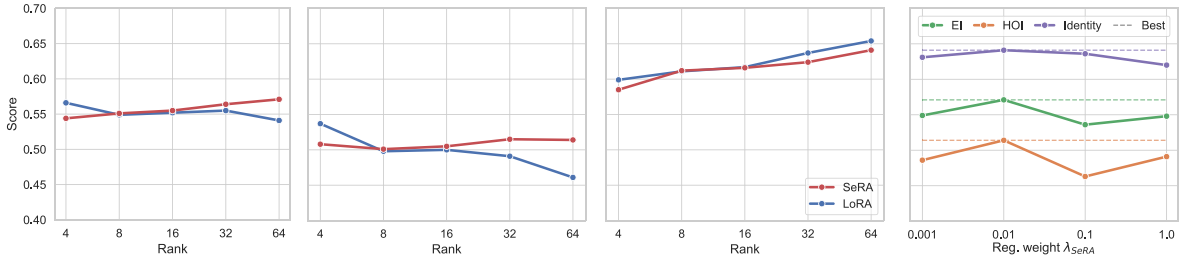
5.4.3. Sensitivity to rank and SeRA regularization strength

We analyze the sensitivity of InteractEdit to the two hyperparameters introduced in Section 3.4: the rank r of the adapter (Eq. 5) and the SeRA uncertainty-loss weight λ_{SeRA} (Eq. 6). For the rank, we sweep $r \in \{4, 8, 16, 32, 64\}$ for *both* vanilla LoRA and SeRA to compare how the ungated and gated adapters scale with capacity; for the regularization, we sweep $\lambda_{\text{SeRA}} \in \{0.001, 0.01, 0.1, 1.0\}$ on SeRA. All variants are evaluated on IEBench setting.

Table 7
Per-action disaggregation of HOI Editability on IEBench-L, sorted by HE (ascending). N is the number of (image, edit) pairs. HE: HOI Editability.

Target verb	N	EI	HE	Identity
groom	6	0.000	0.000	0.207
open	3	0.000	0.000	0.222
fill	4	0.143	0.100	0.248
make	14	0.140	0.114	0.180
stand on	3	0.172	0.133	0.242
feed	24	0.209	0.175	0.259
cut	20	0.213	0.210	0.217
drink with	7	0.230	0.229	0.231
sip	4	0.203	0.250	0.172
wash	26	0.279	0.323	0.245
walk	15	0.297	0.333	0.268
catch	12	0.321	0.417	0.261
smell	6	0.276	0.433	0.203
eat	28	0.282	0.471	0.201
kick	5	0.329	0.520	0.240
kiss	18	0.311	0.567	0.214
eat at	1	0.447	0.600	0.356
lie on	14	0.458	0.614	0.365
hit	10	0.354	0.620	0.247
jump	7	0.502	0.657	0.406
sit on	11	0.497	0.673	0.394
pick up	8	0.279	0.675	0.176
hold	32	0.343	0.750	0.222
sit at	1	0.436	0.800	0.300
carry	4	0.374	0.850	0.240
dribble	12	0.362	0.900	0.226
hug	12	0.431	0.950	0.279
ride	8	0.539	0.975	0.372
throw	10	0.437	0.980	0.281
read	1	0.406	1.000	0.255
Average	326	**0.332**	**0.498**	**0.250**

Rank. The SeRA-vs-LoRA comparison is governed by a single mechanism: as r in Eq. (5) grows, vanilla LoRA spends surplus capacity on memorising source-image structure, while SeRA's gating G closes the surplus directions. At $r = 4$, four rank directions are too few for either adapter to fit the source: LoRA shows slightly higher HE (0.537 vs. SeRA 0.508, Fig. 10b), but Identity is low for both (SeRA 0.585, LoRA 0.599, Fig. 10c), so neither holds an advantage. From $r = 8$ onward the two methods diverge as predicted by Section 3.4: vanilla LoRA's HE drops steadily (0.498 → 0.461, Fig. 10b) and its EI peaks at $r = 32$ before degrading at $r = 64$ (**0.555** → **0.541**, Fig. 10a), while SeRA's gating absorbs the surplus, keeping HE flat (≈ 0.51, Fig. 10b) and growing EI



(a) Editability-Identity (EI) (b) HOI Editability (c) Identity Consistency (d) Regularization strength

Fig. 10. Sensitivity of InteractEdit to rank and SeRA regularization on IEBench. Panels (a)–(c) sweep the adapter rank $r \in \{4, 8, 16, 32, 64\}$ for both vanilla LoRA (blue) and SeRA (red), comparing (a) Editability–Identity (EI), (b) HOI Editability, and (c) Identity Consistency. Panel (d) sweeps the SeRA uncertainty-loss weight $\lambda_{\text{SeRA}} \in \{0.001, 0.01, 0.1, 1.0\}$ on a log axis, with the per-metric best values marked as dashed lines: all three metrics peak at the default $\lambda_{\text{SeRA}} = 0.01$.

Table 8
Per-object disaggregation of HOI Editability on IEBENCH-L, sorted by HE (ascending). Columns as in Table 7.

Object	N	EI	HE	Identity
bottle	6	0.189	0.167	0.219
cup	16	0.207	0.200	0.214
banana	10	0.269	0.240	0.307
cake	48	0.221	0.296	0.177
chair	9	0.292	0.311	0.274
dog	26	0.290	0.369	0.239
broccoli	8	0.289	0.400	0.226
horse	54	0.319	0.411	0.261
cat	20	0.296	0.420	0.228
pizza	8	0.259	0.450	0.182
surfboard	3	0.395	0.467	0.342
dining table	2	0.447	0.700	0.328
skateboard	15	0.510	0.733	0.391
sports ball	60	0.373	0.743	0.249
snowboard	4	0.493	0.750	0.368
bench	12	0.551	0.783	0.425
donut	12	0.292	0.800	0.178
sandwich	3	0.327	0.800	0.206
couch	2	0.408	0.800	0.274
book	6	0.422	0.867	0.279
teddy bear	2	0.390	1.000	0.243
Average	326	**0.332**	**0.498**	**0.250**

monotonically (**0.551** → **0.571**, Fig. 10a). Identity Consistency tracks closely across all ranks (Fig. 10c), so the SeRA-vs-LoRA gap is purely an editability gap.

Regularization. All three metrics peak at the default $\lambda_{\text{SeRA}} = 0.01$ (EI 0.571, HE 0.514, Identity 0.641) in Fig. 10(d) and both weakening ($\lambda_{\text{SeRA}} = 0.001$) and strengthening ($\lambda_{\text{SeRA}} = 0.1$) the uncertainty loss degrade EI and HE. The default $\lambda_{\text{SeRA}} = 0.01$ is therefore the optimal for all three metrics.

5.5. Per-category failure analysis

To identify *where* INTERACTEDIT fails rather than only how often, we disaggregate the IEBENCH-L results by target verb (Table 7) and by object (Table 8).

(i) Verbs. HE varies widely across target verbs. The weakest are *groom* (0.000), *open* (0.000), *fill* (0.100), *make* (0.114), and *stand on* (0.133), while the strongest are *throw* (0.980), *ride* (0.975), *hug* (0.950), and *dribble* (0.900).

(ii) Objects. The weakest objects are small, hand-held, or deformable: *bottle* (0.167), *cup* (0.200), *banana* (0.240), and *cake* (0.296). The strongest are larger and rigid: *bench* (0.783), *snowboard* (0.750), *sports ball* (0.743), and *skateboard* (0.733). This is consistent with the small-object failure mode described in Section 6.1.

5.6. User study

To assess human preference, we conducted a user study comprising a total of 480 individual trials, in which participants were asked to select the best result from six competing methods across 15 unique interaction edits, involving 32 participants in total. To ensure fairness, outputs were randomly shuffled before presentation.

The study included NTI, TurboEdit, and Flux.1 Kontext as the strongest pre-existing interaction editing methods, HOIEdit as an HOI-specific baseline, and Break-A-Scene XL as a disentanglement-based editing baseline. Our method (INTERACTEDIT) was evaluated alongside them.

As shown in Fig. 11, 65.4% of participants preferred INTERACTEDIT. Flux.1 Kontext received 13.8%, Break-A-Scene XL 8.5%, TurboEdit 6.5%, NTI 3.5%, and HOIEdit only 2.3%. The strong preference for our method highlights its ability to balance realism with effective interaction editing, preserving identity while enabling meaningful transformations. More information is detailed in Appendix E.

To verify that our automated metrics: HOI Editability (HE) and the joint Editability–Identity score (EI), are reliable proxies for human perception, we measure their agreement with the user study presented in Appendix E. For each of the six methods, we take the metric score reported in Table 2 and the overall human preference rate (%) in the user study. Fig. 12 plots HE and EI against the human preference rate, together with a linear regression fit. Both metrics show a strong positive correlation with human judgment, with Pearson $r = 0.803$ ($p = 0.054$) for EI and $r = 0.789$ ($p = 0.062$) for HE; rank correlation follows the same trend (Spearman $\rho = 0.600$ for EI and 0.429 for HE). EI consistently aligns more closely with human preference than HE, which is expected by design: HE evaluates only whether the target interaction is realized, while EI jointly captures editing success and identity preservation—the two qualities users implicitly trade off when comparing edited images. The strong, monotonic relationship between EI and human preference confirms that improvements in EI translate into perceivable improvements for human observers, and supports our use of EI as the primary quantitative metric throughout the paper.

6. Conclusion

This paper introduces a novel reference-free interaction editing task and proposes INTERACTEDIT to tackle its challenges. HOI editing remains a formidable problem, requiring interaction modifications while preserving identity. Existing methods struggle to balance HOI editability and identity consistency, as some overfit to the source image and overwrite pretrained interaction priors, while others fail to effectively leverage these priors, leading to suboptimal edits. To address this, INTERACTEDIT employs Selective-Rank Adaptation (SeRA) to regularize inversion and mitigate overfitting, alongside a *Selective Inversion* strategy to preserve target interaction knowledge while adapting

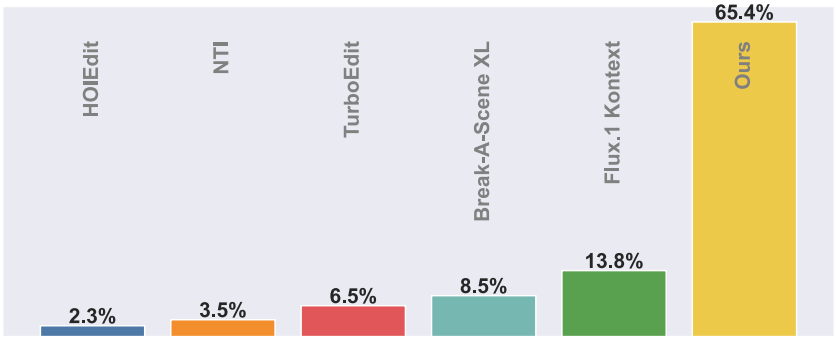


Fig. 11. User preference study results. A total of 480 trials were conducted with 32 participants to assess HOI editing quality. Our method, InteractEdit, was preferred in 65.4% of cases, significantly outperforming the next-best method, Flux.1 Kontext (13.8%).



Fig. 12. Correlation between automated metrics and human preferences. Both HOI Editability (HE) and Editability–Identity (EI) scores show strong positive correlations with human preference rates, with EI exhibiting a slightly stronger relationship (Pearson r = 0.803, p = 0.054) than HE (r = 0.789, p = 0.062). This validates the reliability of our proposed metrics as proxies for human judgment in evaluating HOI editing quality.



(a) Very small object (hit->catch tennis ball)

(b) Rare verb–object composition (*blow*→*make* cake)

(c) Extreme viewpoint change (*eat*→*cut* pizza)

Fig. 13. Representative failure cases of InteractEdit. (a) When the target object occupies only a few pixels (e.g., a tennis ball), the limited spatial resolution of cross-attention maps provides weak supervision for the identity tokens, resulting in degraded object reconstruction. (b) For rare verb–object compositions that are underrepresented in the pretrained diffusion model (e.g., *make a cake*), the reference-free setting cannot reliably synthesise the target interaction. (c) When the target interaction requires very large viewpoint or pose changes (e.g., from a front-facing *eat* to a side-view *cut* pose), the inverted tokens may fail to generalize, causing subtle drift in subject or object appearance.

appearance from the source image. Extensive quantitative and qualitative evaluations demonstrate that our method significantly outperforms state-of-the-art approaches, setting a new benchmark for HOI editing.

6.1. Failure cases and limitations

While InteractEdit advances reference-free HOI editing, we identify four main failure modes, three of which are illustrated in Fig. 13. First, *very small subjects or objects* (Fig. 13a) occupy only a handful of tokens in the cross-attention maps, so the identity tokens receive weak supervision over that region and the inverted appearance is poorly reconstructed in the edited output. Second, *rare verb–object compositions* (Fig. 13b) that are underrepresented in the pretrained diffusion model cannot be reliably synthesised: because our framework is reference-free, it relies entirely on the base model's generative prior for the target interaction, and rare compositions such as *make a cake* fall outside this prior. Third, *extreme non-rigid changes* (Fig. 13c), such as transitions between drastically different viewpoints or body poses (e.g., front-view *eat* to side-view *cut*), push the inverted tokens beyond the appearance distribution they were optimized on, leading to subtle drift in subject or object identity. Fourth, *simultaneous multi-HOI editing* is not supported by our current design: inversion is performed per subject–object–background triplet, so scenes containing multiple interacting pairs must be edited one pair at a time rather than jointly in a single pass. We view these limitations as natural directions for future work, including higher-resolution attention conditioning for small instances, lightweight retrieval-based priors for rare compositions, and a multi-HOI inversion formulation for joint multi-HOI editing.

CRediT authorship contribution statement

Jiun Tian Hoe: Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Weipeng Hu:** Writing – review & editing, Validation, Methodology, Data curation. **Wei Zhou:** Visualization, Software, Data curation. **Chao Xie:** Visualization, Software, Data curation. **Ziwei Wang:** Writing – review & editing, Validation, Supervision, Methodology. **Xudong Jiang:** Writing – review & editing, Supervision, Methodology. **Yap-Peng Tan:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Chee Seng Chan:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Implementation details

Training Setup Our model was trained on NVIDIA 6000 ADA 48GB, with baseline evaluations partially conducted on NVIDIA A100 SXM 80GB. In InteractEdit, both stages utilize Adam optimizer [103] with a batch size of 1 and weight decay of 1e-4. During the inference stage, we set the number of denoising steps to $T = 50$.

Table 9
Comparison between InteractEdit and existing baselines in terms of resource cost. Total time taken is computed as the total required time to complete IEBench, while average time taken is averaged over the number of images generated. † indicates methods requiring pretraining, pretraining time is not included in this table. * "Others" refers to the total time taken for training interaction tokens. ‡ "Others" refers to Induction-2, which requires 100 times. For further details, refer to Appendix A.1.

		Time Taken (s)			Total	Average	GPU Memory (GB)	
Method	Base	Inversion	Editing	Others	(min)	(s)	Inversion	Editing
NTI	SDv1.5	18.81	4.56		85	5.09	11.14	11.15
PnP	SDv2.1	50.16			130	7.81	3.84	7.97
MasaCtrl	SDv1.4	3.68	10.36		174	10.46	6.07	12.48
HOIEdit	SDv1.4	46.93	14.07	81,000	1606	96.38	11.07	20.57
Imagic	SDv1.5	244.98	3.81		472	28.31	18.83	19.93
CDS	SDv1.5	-	34.19		570	34.19	-	10.23
DAC	SDv2.1	330.52	5.07	335.88	799	47.91	16.43	9.71
LEDITS++	SDXL	4.13	4.46		76	4.58	9.45	9.45
InstructPix2Pix †	SDXL	-	74.44		1241	74.44	-	12.36
ReNoise	SDXL Turbo	4.99	0.36		89	5.35	8.38	9.36
TurboEdit	SDXL Turbo	-	1.31		22	1.31	-	23.88
SVDiff	SDv1.5	330.01	22.4		527	31.64	15.89	12.65
Edit Friendly DDPM	SDv1.5	9.07	5.79		101	6.04	6.32	6.95
DDS	SDv2.1	-	9.23		154	9.23	-	7.32
InfEdit	SDv1.5	-	3.92		65	3.92	-	15.58
Break-A-Scene	SDv2.1	274.3	1.42		152	9.10	27.79	4.09
Break-A-Scene XL	SDXL	502.15	5.34		323	19.40	65.84	14.55
Mix-of-Show	SDv1.4	332	1.82		185	11.12	10.10	3.44
DisenDiff	SDv1.5	373	4.4		247	14.83	9.687	10.66
OmniGen	–	-	142.9		2382	142.9	-	10.98
HiDream-E1	HiDream-L1	-	101		1683	101	-	63.23
Flux.1 Kontext	Flux.1	-	59		983	59	-	36.65
InteractEdit	SDXL	466.27	5.92		316	18.98	14.11	14.74

A.1. Resources analysis and computational efficiency

To analyze the resource usage of each method, we evaluate the time and memory consumption, as detailed in Table 9. All experiments are conducted using a single NVIDIA A100 SXM4 80GB GPU, and the analysis is based on the resource usage required to complete IEBENCH. For most image editing methods, the total time taken is the sum of the time spent on source image inversion (28 inversions) and generating the edited images (1000 edits).

However, some methods deviate from this workflow:

- **HOIEdit**: Requires training an interaction token for each action. For IEBench, with 18 different target actions, the training process takes a total of 58,320 s.
- **Imagic**: Performs inversion using both the target prompt and source images, requiring 100 inversions for IEBENCH.
- **DAC**: Involves two types of inversion: Induction-1 (once per source image, 28 inversions) and Induction-2 (per source–target pair, 100 inversions, see `Others` column in Table 9).
- **InstructPix2Pix**: Requires extensive pretraining on a large instruction–image pair dataset. The pretraining dataset[2], contains 313,010 edit instruction–image pairs. InstructPix2Pix was trained for 15,000 steps using 8x NVIDIA A100 GPUs with a batch size of 32.[3]
- **ReNoise**: Performs inversion for each generated image, resulting in 1000 inversions for IEBench.
- **TurboEdit**: Integrates inversion into the editing process.

Due to the inherent properties of latent inversion, some methods (Null-Text Inversion with Prompt-to-Prompt, PnP Diffusion, MasaCtrl, HOIEdit, LEDITS++, and Edit-Friendly DDPM) initialize the editing from a fixed inverted noise rather than random noise, resulting in deterministic editing outcomes with no variation across multiple generations. Nevertheless, to ensure a fair comparison with other methods, we report the time required to generate 10 edits, maintaining consistency with the evaluation protocol.

Our proposed method, InteractEdit, achieves comparable computational efficiency to existing methods while significantly enhancing HOI editability. Unlike HOIEdit, which relies on costly interaction token training, InteractEdit operates in a zero-shot manner without requiring interaction inversion. Furthermore, InteractEdit requires less than 15GB of GPU memory, making it runnable on mainstream GPUs.

A.2. Baseline implementations

To facilitate fair comparisons in the HOI editing task, we re-implemented existing methods as part of our benchmark. This re-implementation aims to standardize evaluation practices and support future research efforts in this area. The re-implemented methods will be made open-source alongside our benchmark. In this section, we provide detailed descriptions of how each method was implemented.

Null-Text Inversion [89] is implemented with Prompt-to-Prompt [104], following the official code.[4] Null-Text Inversion enables real-image inversion, and the editing process is performed with Prompt-to-Prompt. For inversion, we use a null inversion prompt (""), while for editing, the target prompt follows the format: "a photo of [subject] [target interaction] [object]". We use Stable Diffusion v1.5 with a guidance scale of 7.5 and 50 sampling steps.

PnP Diffusion [59] is implemented following the official diffusers code.[5] We adhere to the default hyperparameters: $\tau_f = 0.8, \tau_A = 0.5, \text{timesteps} = 50$ and use Stable Diffusion 2.1. For inversion, we use a null inversion prompt (""), while for editing, the target prompt follows the format: "a photo of [subject] [target interaction] [object]".

MasaCtrl [91] is implemented following the official code.[6] We use Stable Diffusion v1.4 as we are unable to reproduce the method on Stable Diffusion XL. We follow the default hyperparameters: $S = 4, L = 10$. For

[2] `timbrooks/instructpix2pix-clip-filtered`
[3] https://huggingface.co/diffusers/sdxl-instructpix2pix-768.
[4] https://github.com/google/prompt-to-prompt.
[5] https://github.com/MichalGeyer/pnp-diffusers.
[6] https://github.com/TencentARC/MasaCtrl.

inversion, we use a null inversion prompt (“”), while for editing, the target prompt follows the format: “a photo of [subject] [target interaction] [object]”.

HOIEdit [9] is implemented following the official code.[7] We use Stable Diffusion v1.4 and follow the default hyperparameters: $S = 4, L = 10$. Following the paper, for inversion, we use a null inversion prompt (“”), while for editing, the target prompt follows the format: ”[subject] <R> [object]”, which yields the best outcome in preserving identity while accurately modifying the interaction, where <R> is trained interaction token using ReVersion [105]. To train the interaction token for all 25 actions in IEBENCH, we randomly select 10 images as the training images.

Imagic [60] is implemented following the community-provided code.[8] We use Stable Diffusion v1.5 and adopt the default hyperparameters from the community code: $\eta = 1.2$. Although this value differs from the official implementation, we find it produces better results in our experiments. For editing, the target prompt is formatted as: “a photo of [subject] [target interaction] [object]”.

CDS [92] is implemented following the official code.[9] We use Stable Diffusion v1.5 and follow the default hyperparameters: $\ell_{\text{con}} = 3.0, \ell_{\text{DDS}} = 1.0$. For inversion, we use a source prompt that follows the format: “a photo of [subject] [source interaction] [object]”, while for editing, the target prompt follows the format: “a photo of [subject] [target interaction] [object]”.

DAC [65] is implemented following the official code.[10] We use Stable Diffusion v2.1 and follow the default hyperparameters: annealing of 0.8, a guidance scale of 4.0 and 30 inference steps. For abduction 1, which performs the inversion, we use a source prompt that follows the format: “a photo of [subject] [source interaction] [object]”, while for abduction 2, which is related to the editing, we use the target prompt that follows the format: “a photo of [subject] [target interaction] [object]”.

LEDITS++ [64] is implemented following the official code.[11] We use Stable Diffusion XL and follow the default parameters: edit guidance scale of [5.0, 10.0] and edit threshold of [0.9, 0.85] with edit direction of [reverse, forward] for source and target concepts. Since LEDITS++ requires the inverted prompt to be the target concept, we let it be ”[source interaction]”, while the target prompt is ”[target interaction]”.

InstructPix2Pix [58] is implemented following the community-provided code.[12] We use Stable Diffusion XL based pretrained weights `diffusers/sdxl-instructpix2pix-768` and adopt the following hyperparameters: image guidance of 1.5, guidance scale of 3.0 and 100 inference steps. The target prompt is formatted as: “Make [subject] [target interaction] [object]”, as the method requires the prompt to be given as an instruction.

ReNoise [93] is implemented following the official code.[13] We use Stable Diffusion XL Turbo and employ the following hyperparameters: 4 inversion steps, 4 inference steps, 9 renoise steps, $\lambda_{\text{pair}} = 20.0, \lambda_{\text{patch-KL}} = 0.065$. For inversion, we use a source prompt that follows the format: “a photo of [subject] [source interaction] [object]”, while for editing, the target prompt follows the format: “a photo of [subject] [target interaction] [object]”.

TurboEdit [94] is implemented following the official code.[14] We use Stable Diffusion XL Turbo and employ the following hyperparameters: pseudo-guidance scale $w = 1.5$ and 4 denoising steps. For inversion, we use a source prompt that follows the format: “a photo of [subject] [source interaction] [object]”, while for editing, the target prompt follows the format: “a photo of [subject] [target interaction] [object]”.

SVDiff [44] is implemented following the community-provided code.[15] We use Stable Diffusion v1.5. For inversion, we use a source prompt that follows the format: “a photo of [subject] [source interaction] [object]”, while for editing, the target prompt follows the format: “a photo of [subject] [target interaction] [object]”.

Edit Friendly DDPM [95] is implemented following the official code.[16] We use Stable Diffusion v1.5 and the following default hyperparameters: strength $= 15, T_{\text{skip}} = 36, \eta = 1$ and 100 inference steps. For inversion, we use a source prompt that follows the format: “a photo of [subject] [source interaction] [object]”, while for editing, the target prompt follows the format: “a photo of [subject] [target interaction] [object]”.

DDS [96] is implemented following the official code.[17] We use Stable Diffusion v2.1 and use the default hyperparameters. For inversion, we a source prompt that follows the format: “a photo of [subject] [source interaction] [object]”, while for editing, the target prompt follows the format: “a photo of [subject] [target interaction] [object]”.

InfEdit [97] is implemented following the official code.[18] Following the paper, we use Stable Diffusion v1.5 (`SimianLuo/LCM_Dreamshaper_v7`) and use the following default hyperparameters: source guidance scale of 1, target guidance scale of 2, cross replace steps of 0.3, self replace steps of 0.3, target blend threshold of 0.3, source blend threshold of 0.3, and inference steps of 15. For inversion, we use a source prompt that follows the format: “a photo of [subject] [source interaction] [object]”, while for editing, the target prompt follows the format: “a photo of [subject] [target interaction] [object]”.

Break-A-Scene [68] is implemented following the official code[19]. We use Stable Diffusion v2.1 and the following default hyperparameters: $\lambda_{\text{attn}} = 0.01$. We fine-tune for 400 steps in stage 1 and 800 steps in stage 2 with the default learning rate of 5e-4 and 2e-6, which we found to work well empirically. For fair comparison, we re-implement **Break-A-Scene XL** which is based on Stable Diffusion XL (`SG161222/RealVisXL_V4.0`). We use $\lambda_{\text{attn}} = 0.01$ and fine-tune for 1000 steps in stage 1 and 200 steps in stage 2, with a learning rate of 3e-4 and 2e-6, respectively, which we found to work well empirically.

Mix-of-Show [74] is implemented using the official code[20] and default settings. For inversion, the subject and object are each represented by two tokens, which are processed independently and then fused via gradient fusion. For editing, the prompt is “a photo of [subject tokens] [target interaction] [object tokens]”.

DisenDiff [98] is implemented following the official code[21] and default settings. During inversion, we use the default prompt format: “<new1> [subject] and <new2> [object]”, while during editing, we use “a photo of <new1> [subject] [target interaction] <new2> [object]”.

OmniGen [99] is implemented using diffusers library. There are two prompting strategies: single image and part-splitting. Single image prompt uses only one source image and follows the format: “Edit <img><|image_1|></img> to change the interaction of the [subject] to [target interaction] with the [object].” The part-splitting approach splits the image into subject, object and background, and uses the following prompt: “A [subject] <img><|image_1|></img> [target interaction] with the [object] <img><|image_2|></img> at

[7] https://github.com/Kenneth-Wong/hoiedit.
[8] https://github.com/huggingface/diffusers/blob/main/examples/community/imagic_stable_diffusion.py.
[9] https://github.com/HyelinNAM/ContrastiveDenoisingScore.
[10] https://github.com/xuesong39/DAC.
[11] https://github.com/huggingface/diffusers/tree/main/src/diffusers/pipelines/ledits_pp.
[12] https://github.com/huggingface/diffusers/tree/main/examples/instruct_pix2pix.
[13] https://github.com/garibida/ReNoise-Inversion.
[14] https://github.com/GiilDe/turbo-edit.
[15] https://github.com/mkshing/svdiff-pytorch.
[16] https://github.com/inbarhub/DDPM_inversion.
[17] https://github.com/google/prompt-to-prompt/blob/main/DDS_zeroshot.ipynb.
[18] https://github.com/sled-group/InfEdit.
[19] https://github.com/google/break-a-scene.
[20] https://github.com/TencentARC/Mix-of-Show.
[21] https://github.com/Monalissaa/DisenDiff.

<img><|image_3|></img>." We observe that part-splitting offers better interaction editability, and report scores based on this setting.

HiDream-E1 [100] is implemented following the official code[22]. Following the default setting, input image is resized to 768 × 768 and we adopt the prompt format:"the [subject] is now [target interaction] [object]". This prompt is then further refined using OpenAI GPT-4o, conditioned on the source image, using the prompt engineering protocol provided in the official code.

Flux.1 Kontext [72] is implemented using the diffusers library and model weights[23]. We follow the default hyperparameters: guidance scale of 2.5 and prompt format of "the [subject] is now [target interaction] [object]".

FireFlow [101] is implemented following the official code[24]. Following the default settings, we adopt the source prompt format:"a photo [subject] [source interaction] [object]" and the target prompt format of "a photo of [subject] [target interaction] [object]". We use guidance scale of 3, `flux-dev` base model, `add_q` editing strategy, `fireflow` sampling strategy, 8 inference steps, `end_layer_index` at 17, and use default settings for editing strategy.

Appendix B. CLIP-T sensitivity analysis

To systematically evaluate the sensitivity of CLIP-T to actions (verbs) versus objects (nouns), we conducted an analysis on 1911 images sampled from the HICO-DET test set. To avoid confounding factors, we filtered the dataset to include only images containing a single HOI instance. For each image, we generated mismatched prompts in two ways: (i) by replacing the ground-truth action with all other possible verbs applicable to the same object (verb mismatch), and (ii) by replacing the ground-truth object with all other possible nouns for the same action (noun mismatch). This yielded a total of 50,239 mismatched prompts: 36,872 object mismatches and 13,367 action mismatches. We then computed the cosine similarity between each mismatched prompt and the image embedding using CLIP, and compared it against the similarity score for the correct image–prompt pair. The score difference reflects CLIP-T's ability to discriminate mismatches. Our analysis shows that verb mismatches produce extremely small margins clustered near zero, whereas noun mismatches yield much broader margins, confirming that CLIP is primarily sensitive to objects but weak at distinguishing actions.

Appendix C. Theoretical proofs

C.1. Capacity control and prior preservation

Lemma 1 (Logit-Subspace Invariance under Selective Inversion). *Let the pretrained query matrix be $W_Q^{(0)} \in \mathbb{R}^{d\times d}$ and define the query-feature map $\Phi(X) = XW_Q^{(0)}$. If, during inversion, W_Q is kept fixed $(W_Q \equiv W_Q^{(0)})$ while W_K is updated arbitrarily, then for every input X the attention-score matrix*

$$S(X) = \frac{1}{\sqrt{d}}\Phi(X)(XW_K)^\top$$

has its row-space *contained in* $\mathrm{row}\big(\Phi(X)\big)$*, which is exactly the row-space learned by the pretrained model. Consequently, all edited images lie in the same interaction-prior subspace as the original network.*

Proof. Because $\Phi(X) = XW_Q^{(0)}$ is fixed, every row of $S(X)$ is a linear combination of the rows of $\Phi(X)$ (scalar weights given by the corresponding columns of XW_K), hence $\mathrm{row}(S(X)) \subseteq \mathrm{row}(\Phi(X))$, independent of the choice of W_K. □

Appendix D. More qualitative results

In Figs. 15 and 16, we provide a more visual comparison on same source images as in Fig. 6 for methods that are not included in the main paper. Fig. 17 extends the comparison to additional source images. Compared to other methods, our method successfully enables interactive edits while maintaining the identity.

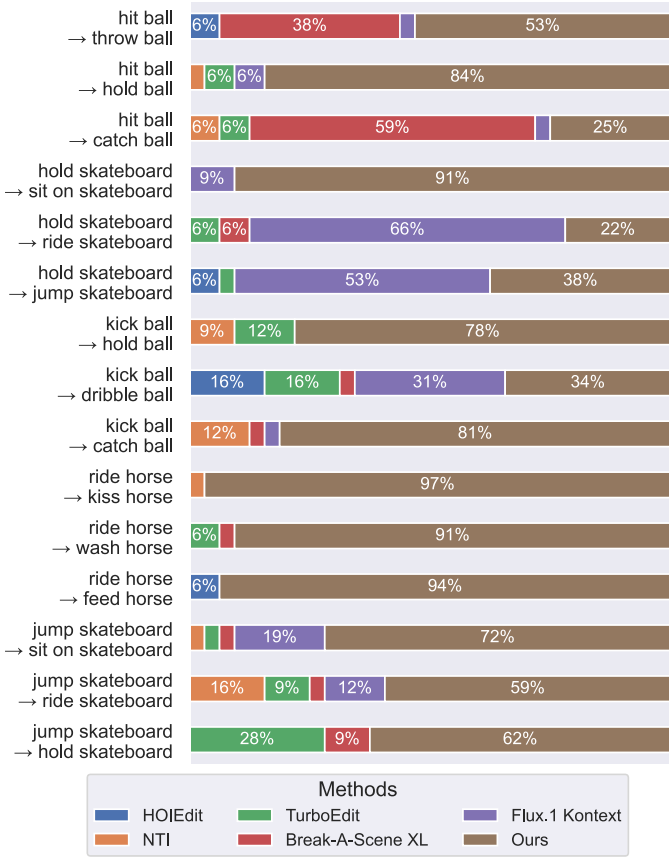


Fig. 14. User preferences for each interaction edit.

[22] https://github.com/HiDream-ai/HiDream-E1.

[23] `black-forest-labs/FLUX.1-Kontext-dev`.

[24] https://github.com/HolmesShuan/FireFlow-Fast-Inversion-of-Rectified-Flow-for-Image-Semantic-Editing.git.

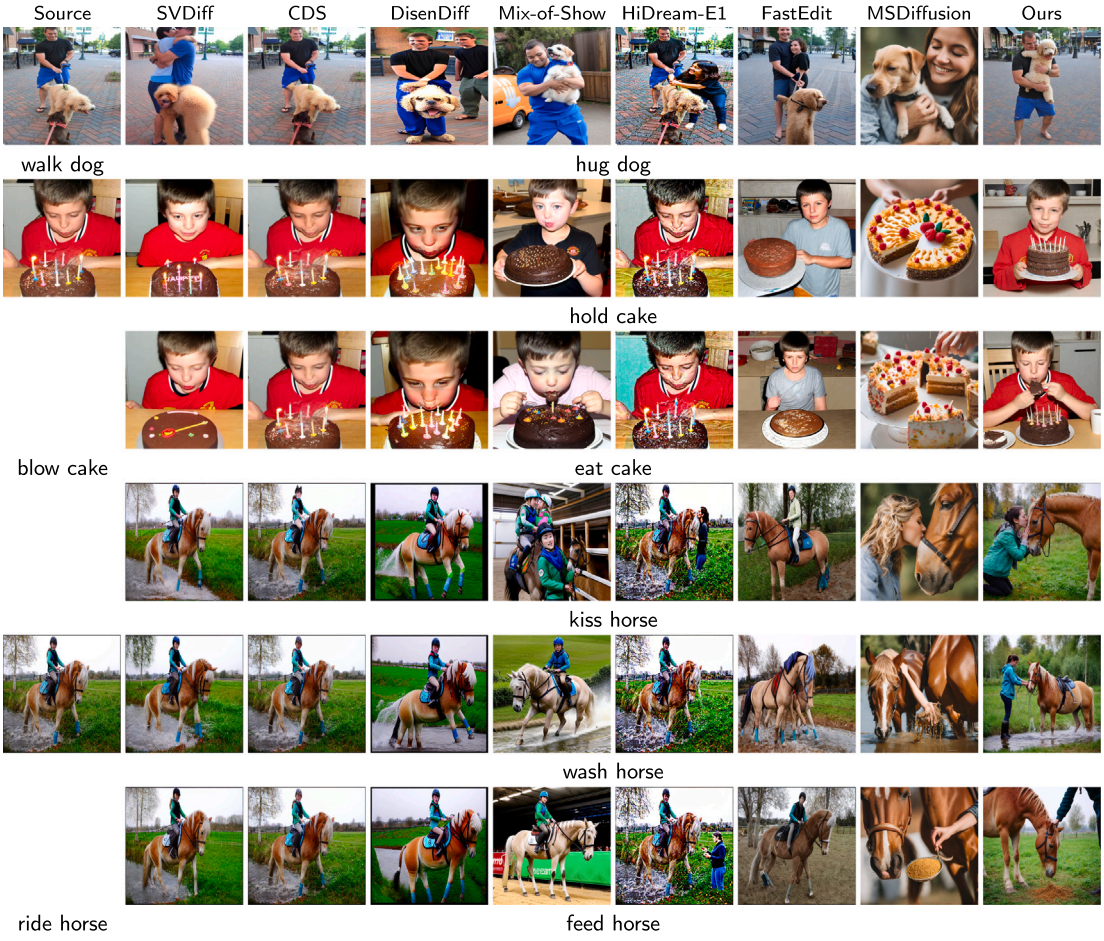


Fig. 15. Qualitative comparison with additional existing baselines. The source column shows the source image and its original interaction.

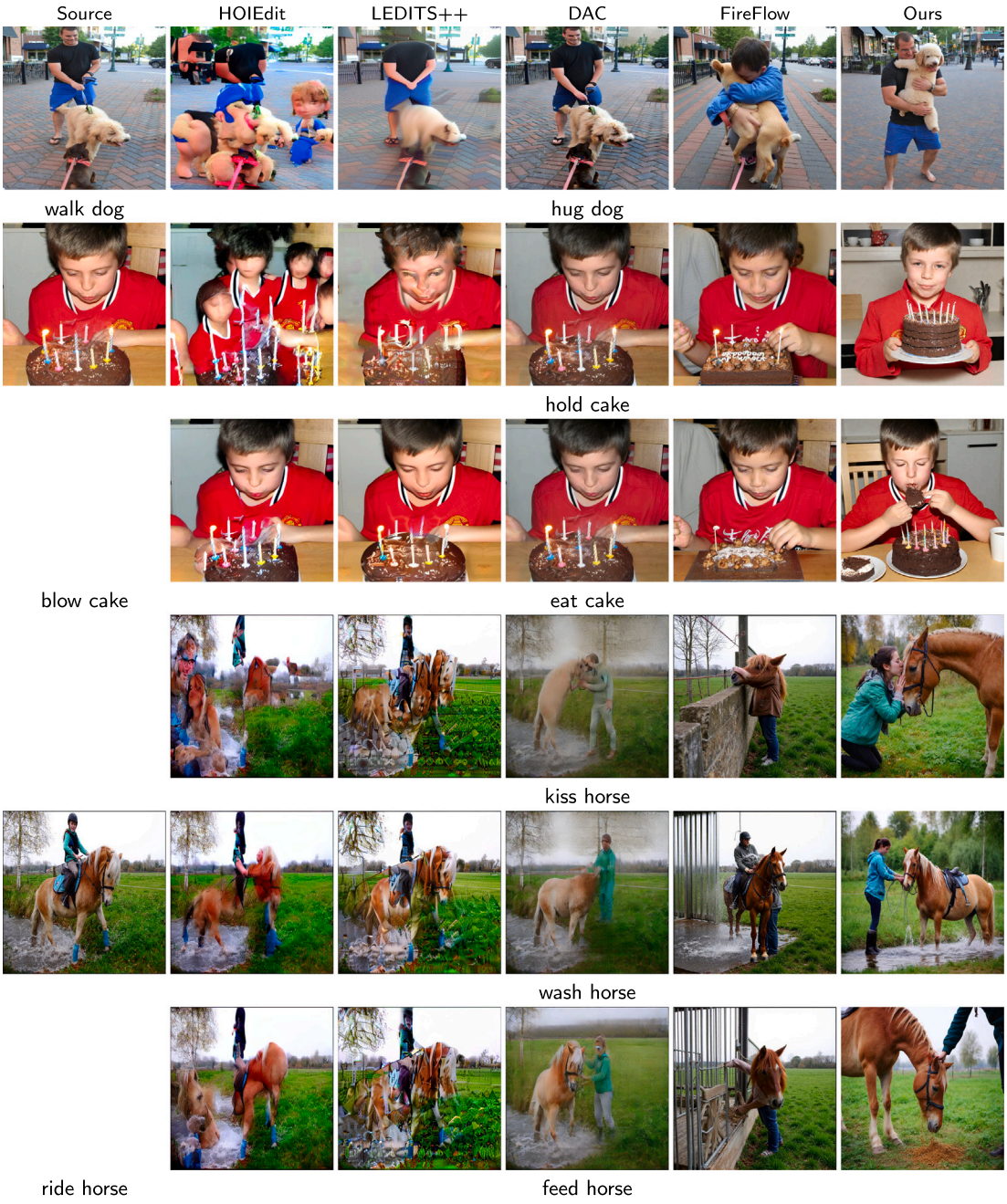


Fig. 16. Qualitative comparison with additional existing baselines.



Fig. 17. More qualitative comparisons. Our method demonstrates the best HOI editability.

Appendix E. User study details

To assess user preferences, we conducted an online survey where participants were instructed to evaluate the edits based on three criteria: (1) adherence to the target interaction—how well the edit reflects the intended action, (2) identity consistency—whether the subject and object retain their original appearance, and (3) realism—how natural and visually plausible the edit appears.

In our user study, participants evaluated edits ranging from subtle action transitions to significant pose and structural modifications. Complex edits such as *jump → hold/ride/sit on skateboard* require adjusting both the subject's pose and the spatial relationship with the object, making them particularly challenging for existing methods. Similarly, *ride → feed/wash/kiss horse* involves close-contact interactions that demand precise alterations to the subject's positioning while preserving identity consistency.

Fig. 14 provides insights into user preferences for different interaction edits. Across 12 out of 15 interaction edits, our method is the most preferred. In some more challenging edits, such as *hold skateboard → sit on skateboard* and *ride horse → kiss/wash/feed horse*, our method significantly outperforms others, due to its ability to modify interactions while preserving identity.

Edits involving a sports ball, such as *kick → catch/dribble/hold ball* assess a model's ability to modify hand and foot placement while maintaining a natural transition between actions. Similarly, *hold → jump/ride/sit on skateboard* requires adapting the subject's posture while ensuring the skateboard remains a consistent object in the scene. Finally, hand–object interaction edits like *hit → catch/hold/hit ball* evaluate whether the model can effectively modify the action without introducing artifacts or unnatural grasps on the object. These diverse edits allow us to assess how well different methods handle a range of interaction modifications, from subtle changes to full-body pose adjustments.

Data availability

Data will be made available upon request.

References

[1] R. Morais, V. Le, S. Venkatesh, T. Tran, Learning asynchronous and sparse human-object interaction in videos, in: CVPR, 2021, pp. 16041–16050.

[2] Y. Yang, W. Zhai, H. Luo, Y. Cao, Z.-J. Zha, Lemon: learning 3D human-object interaction relation from 2D images, in: CVPR, 2024, pp. 16284–16295.

[3] J. Gao, Z. Wang, Z. Xiao, J. Wang, T. Wang, J. Cao, X. Hu, S. Liu, J. Dai, J. Pang, Coohoi: learning cooperative human-object interaction with manipulated object dynamics, NeurIPS 37 (2024) 79741–79763.

[4] L. Li, W. Wang, Y. Yang, Human-object interaction detection collaborated with large relation-driven diffusion models, arXiv preprint arXiv:2410.20155, 2024.

[5] J. Luo, W. Ren, W. Jiang, X. Chen, Q. Wang, Z. Han, H. Liu, Discovering syntactic interaction clues for human-object interaction detection, in: CVPR, 2024, pp. 28212–28222.

[6] J.T. Hoe, X. Jiang, C.S. Chan, Y.-P. Tan, W. Hu, InteractDiffusion: interaction control in text-to-image diffusion models, in: CVPR, 2024, pp. 6180–6189.

[7] X. Hu, H. Wang, J.E. Lenssen, B. Schiele, PersonaHOI: effortlessly improving personalized face with human-object interaction generation, arXiv preprint arXiv:2501.05823, 2025.

[8] X. Shuai, H. Ding, X. Ma, R. Tu, Y.-G. Jiang, D. Tao, A survey of multimodal-guided image editing with text-to-image diffusion models, arXiv preprint arXiv:2406.14555, 2024.

[9] T. Xu, W. Wang, A. Zhong, HOIEdit: human–object interaction editing with text-to-image diffusion model, Vis. Comput. (2025) 1–13.

[10] J.T. Hoe, W. Hu, W. Zhou, C. Xie, Z. Wang, C.S. Chan, X. Jiang, Y.-P. Tan, InteractEdit: zero-shot editing of human-object interactions in images, Preprint arXiv:2503.09130, 2025.

[11] F.Z. Zhang, Y. Yuan, D. Campbell, Z. Zhong, S. Gould, Exploring predicate visual context in detecting human–object interactions, in: ICCV, 2023, pp. 10411–10421.

[12] S. Fang, Z. Lin, K. Yan, J. Li, X. Lin, R. Ji, Hodn: disentangling human-object feature for hoi detection, IEEE Trans. Multimed. 26 (2023) 3125–3136.

[13] T. Luo, Q. Yuan, B. Zhu, S. Guan, R. Yang, J.S. Smith, E. Gee Lim, Simple yet effective: an explicit query-based relation learner for human–object-interaction detection, Neurocomputing 649 (2025) 130709, https://doi.org/10.1016/j.neucom.2025.130709, https://www.sciencedirect.com/science/article/pii/S0925231225013815.

[14] T. Luo, Q. Yuan, B. Zhu, S. Guan, R. Yang, J.S. Smith, E.G. Lim, Exploring interaction concepts for human–object-interaction detection via global- and local-scale enhancing, Neurocomputing 651 (2025) 130882, https://doi.org/10.1016/j.neucom.2025.130882, https://www.sciencedirect.com/science/article/pii/S0925231225015541.

[15] M. Zou, S. Liu, Q. Zeng, X. Zhang, G. Yuan, X. Hao, J. Liu, W. Zhou, Pose-guided multi-cue explicit query construction for disambiguating human–object interactions, IEEE Trans. Circuits Syst. Video Technol. 36 (7) (2026) 10794–10809, https://doi.org/10.1109/TCSVT.2026.3667102

[16] X. Han, X. Zhang, G. Niu, M. Zhou, Z. Pan, Soft-label guided multi-granularity prompts learning for human-object interaction detection, Appl. Soft Comput. 192 (2026) 114765, https://doi.org/10.1016/j.asoc.2026.114765, https://www.sciencedirect.com/science/article/pii/S1568494626002139.

[17] X. Zhang, X. Li, M. Zhou, M. Gan, C.L. Philip Chen, ASCFormer: an adaptive structure-aware cascaded transformer for 3D object detection, IEEE Trans. Circuits Syst. Video Technol. 36 (2) (2026) 2218–2232, https://doi.org/10.1109/TCSVT.2025.3612592

[18] J.-Y. Jiang-Lin, K.-Y. Huang, L. Lo, Y.-N. Huang, T. Lin, J.-C. Wu, H.-H. Shuai, W.-H. Cheng, ReCorD: reasoning and correcting diffusion for HOI generation, in: ACM-Mm, 2024, pp. 9465–9474.

[19] M. Ge, X. Jia, T. Isobe, X. Li, Q. Wang, J. Mu, D. Zhou, L. Wang, H. Lu, L. Tian, et al., Customizing text-to-image generation with inverted interaction, in: ACM-Mm, 2024, pp. 10901–10909.

[20] S. Cha, K. Lee, Y.-C. Kim, H. Oh, D.-J. Kim, VerbDiff: text-only diffusion models with enhanced interaction awareness, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 8041–8050.

[21] A. Taghipour, M. Ghahremani, M. Bennamoun, A.M. Rekavandi, H. Laga, F. Boussaid, Box it to bind it: unified layout control and attribute binding in text-to-image diffusion models, IEEE Trans. Multimed. 27 (2025).

[22] J. Li, W. Hu, X. Wang, J.T. Hoe, P. Hu, X. Jiang, Y.-P. Tan, Adaptive memory refinement and perception enhancement for exo-to-ego video generation, Neurocomputing 660 (2026) 131917, https://doi.org/10.1016/j.neucom.2025.131917, https://www.sciencedirect.com/science/article/pii/S0925231225025895.

[23] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, NeurIPS 33 (2020) 6840–6851.

[24] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, B. Poole, Score-based generative modeling through stochastic differential equations, in: ICLR, 2021, pp. 1–36, https://openreview.net/forum?id=PxTIG12RRHS.

[25] P. Dhariwal, A. Nichol, Diffusion models beat GANs on image synthesis, NeurIPS 34 (2021) 8780–8794.

[26] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, M. Chen, Glide: towards photorealistic image generation and editing with text-guided diffusion models, arXiv preprint arXiv:2112.10741, 2021.

[27] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, arXiv preprint arXiv:2204.06125, 2022.

[28] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, R. Rombach, SDXL: improving latent diffusion models for high-resolution image synthesis, arXiv preprint arXiv:2307.01952, 2023.

[29] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, Y.J. Lee, Gligen: open-set grounded text-to-image generation, in: CVPR, 2023, pp. 22511–22521.

[30] G. Zheng, X. Zhou, X. Li, Z. Qi, Y. Shan, X. Li, LayoutDiffusion: controllable diffusion model for layout-to-image generation, in: CVPR, 2023, pp. 22490–22499.

[31] J. Xie, Y. Li, Y. Huang, H. Liu, W. Zhang, Y. Zheng, M.Z. Shou, Boxdiff: text-to-image synthesis with training-free box-constrained diffusion, in: ICCV, 2023, pp. 7452–7461.

[32] Q. Phung, S. Ge, J.-B. Huang, Grounded text-to-image synthesis with attention refocusing, in: CVPR, 2024, pp. 7932–7942.

[33] J. Zheng, N. Xu, W. Li, J. Jiang, X. Zhang, Semantic-spatial attention for refined object placement in text-to-image synthesis, IEEE Trans. Multimed. 27 (2025).

[34] H. Ye, J. Zhang, S. Liu, X. Han, W. Yang, IP-adapter: text compatible image prompt adapter for text-to-image diffusion models, arXiv preprint arXiv:2308.06721, 2023.

[35] L. Zhang, A. Rao, M. Agrawala, Adding conditional control to text-to-image diffusion models, in: ICCV, 2023, pp. 3836–3847.

[36] S. Mo, F. Mu, K.H. Lin, Y. Liu, B. Guan, Y. Li, B. Zhou, FreeControl: training-free spatial control of any text-to-image diffusion model with any condition, arXiv preprint arXiv:2312.07536, 2023.

[37] W. Liu, X. Zhong, J. Hou, S. Li, H. Huang, W. Zhou, Y. Fang, Integrating large foundation models into multimodal named entity recognition with evidential fusion, Neurocomputing 652 (2025) 131015, https://doi.org/10.1016/j.neucom.2025.131015, https://www.sciencedirect.com/science/article/pii/S092523122501687X.

[38] L. Zhang, T. Li, B. Lu, J. Fang, D. Zheng, W. Zhou, W. Liu, F. Lv, Rethinking the effect of unimodal labels in multimodal sentiment analysis, ACM Trans. Multimed. Comput. Commun. Appl. 22 (7) (Jun 2026), https://doi.org/10.1145/3796718

[39] Y. Yan, J. Xu, B.-W. Zhang, Z. Zhang, G. Liu, ReVisE: reflective visual editing for multi-instruction image manipulation, Neurocomputing 669 (2026) 132424, https://doi.org/10.1016/j.neucom.2025.132424, https://www.sciencedirect.com/science/article/pii/S0925231225030966.

[40] Z. Yu, J. Jin, J. Zhao, Z. Fu, J. Yang, TtfDiffusion: training-free and text-free image editing in diffusion models with structural and semantic disentanglement, Neurocomputing 619 (2025) 129159, https://doi.org/10.1016/j.neucom.2024.129159, https://www.sciencedirect.com/science/article/pii/S0925231224019301.

[41] J. Shi, W. Xiong, Z. Lin, H.J. Jung, Instantbooth: personalized text-to-image generation without test-time finetuning, in: CVPR, 2024, pp. 8543–8552.

[42] D. Li, J. Li, S. Hoi, BLIP-diffusion: pre-trained subject representation for controllable text-to-image generation and editing, NeurIPS 36 (2024).

[43] Z. Liu, Y. Zhang, Y. Shen, K. Zheng, K. Zhu, R. Feng, Y. Liu, D. Zhao, J. Zhou, Y. Cao, Cones 2: customizable image synthesis with multiple subjects, in: NeurIPS, 2023, pp. 57500–57519.

[44] L. Han, Y. Li, H. Zhang, P. Milanfar, D. Metaxas, F. Yang, Svdiff: compact parameter space for diffusion fine-tuning, in: ICCV, 2023, pp. 7323–7334.

[45] X. Wang, S. Fu, Q. Huang, W. He, H. Jiang, MS-diffusion: multi-subject zero-shot image personalization with layout guidance, arXiv preprint arXiv:2406.07209, 2024.

[46] J. Ma, J. Liang, C. Chen, H. Lu, Subject-diffusion: open domain personalized text-to-image generation without test-time fine-tuning, in: ACM SIGGRAPH 2024 Conference Papers, 2024, pp. 1–12.

[47] K.W. Ng, X. Zhu, Y.-Z. Song, T. Xiang, PartCraft: crafting creative objects by parts, in: ECCV, 2024, pp. 420–437.

[48] Q. Wang, X. Bai, H. Wang, Z. Qin, A. Chen, H. Li, X. Tang, Y. Hu, Instantid: zero-shot identity-preserving generation in seconds, arXiv preprint arXiv:2401.07519, 2024.

[49] C. Shi, X. Wang, S. Shi, X. Wang, M. Zhu, N. Wang, X. Gao, Foodfusion: a novel approach for food image composition via diffusion models, IEEE Trans. Multimed. 27 (2025) 8644–8655.

[50] Y. Jiang, Q. Liu, D. Chen, L. Yuan, Y. Fu, AnimeDiff: customized image generation of anime characters using diffusion model, IEEE Trans. Multimed. 26 (2024).

[51] D. Epstein, A. Jabri, B. Poole, A. Efros, A. Holynski, Diffusion self-guidance for controllable image generation, NeurIPS 36 (2023) 16222–16239.

[52] C. Mou, X. Wang, J. Song, Y. Shan, J. Zhang, Diffeditor: boosting accuracy and flexibility on diffusion-based image editing, in: CVPR, 2024, pp. 8488–8497.

[53] Y. Shi, C. Xue, J.H. Liew, J. Pan, H. Yan, W. Zhang, V.Y. Tan, S. Bai, Dragdiffusion: harnessing diffusion models for interactive point-based image editing, in: CVPR, 2024, pp. 8839–8849.

[54] C. Mou, X. Wang, J. Song, Y. Shan, J. Zhang, Dragondiffusion: enabling drag-style manipulation on diffusion models, arXiv preprint arXiv:2307.02421, 2023.

[55] C. Yang, C. Kang, K. Kong, H. Oh, S.-J. Kang, Person in place: generating associative skeleton-guidance maps for human-object interaction image editing, in: CVPR, 2024, pp. 8164–8175.

[56] H. Liu, C. Xu, Y. Yang, L. Zeng, S. He, Drag your noise: interactive point-based editing via diffusion semantic propagation, in: CVPR, 2024, pp. 6743–6752.

[57] S. Wang, C. Saharia, C. Montgomery, J. Pont-Tuset, S. Noy, S. Pellegrini, Y. Onoe, S. Laszlo, D.J. Fleet, R. Soricut, et al., Imagen editor and editbench: advancing and evaluating text-guided image inpainting, in: CVPR, 2023, pp. 18359–18369.

[58] T. Brooks, A. Holynski, A.A. Efros, InstructPix2Pix: learning to follow image editing instructions, in: CVPR, 2023, pp. 18392–18402.

[59] N. Tumanyan, M. Geyer, S. Bagon, T. Dekel, Plug-and-play diffusion features for text-driven image-to-image translation, in: CVPR, 2023, pp. 1921–1930.

[60] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, M. Irani, Imagic: text-based real image editing with diffusion models, in: CVPR, 2023, pp. 6007–6017.

[61] Z. Geng, B. Yang, T. Hang, C. Li, S. Gu, T. Zhang, J. Bao, Z. Zhang, H. Li, H. Hu, et al., Instructdiffusion: a generalist modeling interface for vision tasks, in: CVPR, 2024, pp. 12709–12720.
[62] H. Hu, K.C. Chan, Y.-C. Su, W. Chen, Y. Li, K. Sohn, Y. Zhao, X. Ben, B. Gong, W. Cohen, et al., Instruct-imagen: image generation with multi-modal instruction, in: CVPR, 2024, pp. 4754–4763.
[63] L. Tsaban, A. Passos, Ledits: real image editing with DDPM inversion and semantic guidance, arXiv preprint arXiv:2307.00522, 2023.
[64] M. Brack, F. Friedrich, K. Kornmeier, L. Tsaban, P. Schramowski, K. Kersting, A. Passos, Ledits++: limitless image editing using text-to-image models, in: CVPR, 2024, pp. 8861–8870.
[65] X. Song, J. Cui, H. Zhang, J. Chen, R. Hong, Y.-G. Jiang, Doubly abductive counterfactual inference for text-based image editing, in: CVPR, 2024, pp. 9162–9171.
[66] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, S. Ermon, SDEdit: guided image synthesis and editing with stochastic differential equations, in: ICLR, 2022, pp. 1–33.
[67] Z. Jiang, C. Mao, Y. Pan, Z. Han, J. Zhang, Scedit: efficient and controllable image diffusion generation via skip connection editing, in: CVPR, 2024, pp. 8995–9004.
[68] O. Avrahami, K. Aberman, O. Fried, D. Cohen-Or, D. Lischinski, Break-a-scene: extracting multiple concepts from a single image, in: SIGGRAPH Asia 2023 Conference Papers, 2023, pp. 1–12.
[69] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, K. Aberman, Dreambooth: fine tuning text-to-image diffusion models for subject-driven generation, in: CVPR, 2023, pp. 22500–22510.
[70] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A.H. Bermano, G. Chechik, D. Cohen-Or, An image is worth one word: personalizing text-to-image generation using textual inversion, arXiv preprint arXiv:2208.01618, 2022.
[71] C. Wei, Z. Xiong, W. Ren, X. Du, G. Zhang, W. Chen, Omniedit: building image editing generalist models through specialist supervision, in: The Thirteenth International Conference on Learning Representations, 2024, pp. 1–21.
[72] B.F. Labs, S. Batifol, A. Blattmann, F. Boesel, S. Consul, C. Diagne, T. Dockhorn, J. English, Z. English, P. Esser, et al., FLUX. 1 kontext: flow matching for in-context image generation and editing in latent space, arXiv preprint arXiv:2506.15742, 2025.
[73] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: low-rank adaptation of large language models, in: ICLR, 2022, pp. 1–13, https://openreview.net/forum?id=nZeVKeeFYf9.
[74] Y. Gu, X. Wang, J.Z. Wu, Y. Shi, Y. Chen, Z. Fan, W. Xiao, R. Zhao, S. Chang, W. Wu, et al., Mix-of-show: decentralized low-rank adaptation for multi-concept customization of diffusion models, NeurIPS 36 (2024).
[75] Y. Frenkel, Y. Vinker, A. Shamir, D. Cohen-Or, Implicit style-content separation using b-lora, in: European Conference on Computer Vision, Springer, 2024, pp. 181–198.
[76] J. Lei, L. Li, C. Wang, J. Xiao, L. Chen, Seeing beyond classes: zero-shot grounded situation recognition via language explainer, in: ACM-Mm, 2024, pp. 1602–1611.
[77] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, J. Deng, Learning to detect human-object interactions, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 381–389.
[78] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[79] L. Momeni, M. Caron, A. Nagrani, A. Zisserman, C. Schmid, Verbs in action: improving verb understanding in video-language models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15579–15591.
[80] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, C. Ross, Winoground: probing vision and language models for visio-linguistic compositionality, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5238–5248.
[81] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al., Grounding DINO: marrying DINO with grounded pre-training for open-set object detection, arXiv preprint arXiv:2303.05499, 2023.
[82] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, et al., Segment anything, in: ICCV, 2023, pp. 4015–4026.
[83] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., DINOv2: learning robust visual features without supervision, arXiv preprint arXiv:2304.07193, 2023.
[84] C.J.V. Rijsbergen, Information Retrieval, second ed, Butterworth-Heinemann, USA, 1979.
[85] X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, H. Li, Human preference score v2: a solid benchmark for evaluating human preferences of text-to-image synthesis, arXiv preprint arXiv:2306.09341, 2023.
[86] H. Wang, J. Liu, H. Tan, J. Lou, X. Liu, W. Zhou, H. Liu, Blind image quality assessment via adaptive graph attention, IEEE Trans. Circuits Syst. Video Technol. 34 (10) (2024) 10299–10309, https://doi.org/10.1109/TCSVT.2024.3405789
[87] Y. Li, X. Yang, G. Yue, J. Fu, Q. Jiang, X. Jia, P.L. Rosin, H. Liu, W. Zhou, Perception-oriented bidirectional attention network for image super-resolution quality assessment, IEEE Trans. Image Process. 34 (2025) 7728–7743, https://doi.org/10.1109/TIP.2025.3633145
[88] Y. Li, X. Yang, J. Fu, G. Yue, W. Zhou, Deep bi-directional attention network for image super-resolution quality assessment, in: 2024 IEEE International Conference on Multimedia and Expo (ICME), 2024, pp. 1–6, https://doi.org/10.1109/ICME57554.2024.10687430
[89] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, D. Cohen-Or, Null-text inversion for editing real images using guided diffusion models, in: CVPR, 2023, pp. 6038–6047.
[90] X. Ju, A. Zeng, Y. Bian, S. Liu, Q. Xu, Pnp inversion: boosting diffusion-based editing with 3 lines of code, in: ICLR, 2024, pp. 1–27.
[91] M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, Y. Zheng, Masactrl: tuning-free mutual self-attention control for consistent image synthesis and editing, in: ICCV, 2023, pp. 22560–22570.
[92] H. Nam, G. Kwon, G.Y. Park, J.C. Ye, Contrastive denoising score for text-guided latent diffusion image editing, in: CVPR, 2024, pp. 9192–9201.
[93] D. Garibi, O. Patashnik, A. Voynov, H. Averbuch-Elor, D. Cohen-Or, ReNoise: real image inversion through iterative noising, arXiv:2403.14602, 2024.
[94] G. Deutch, R. Gal, D. Garibi, O. Patashnik, D. Cohen-Or, Turboedit: text-based image editing using few-step diffusion models, in: SIGGRAPH Asia 2024 Conference Papers, 2024, pp. 1–12.
[95] I. Huberman-Spiegelglas, V. Kulikov, T. Michaeli, An edit friendly DDPM noise space: inversion and manipulations, in: CVPR, 2024, pp. 12469–12478.
[96] A. Hertz, K. Aberman, D. Cohen-Or, Delta denoising score, in: ICCV, 2023, pp. 2328–2337.
[97] S. Xu, Y. Huang, J. Pan, Z. Ma, J. Chai, Inversion-free image editing with natural language, in: CVPR, 2024, pp. 2328–2337.
[98] Y. Zhang, M. Yang, Q. Zhou, Z. Wang, Attention calibration for disentangled text-to-image personalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.
[99] S. Xiao, Y. Wang, J. Zhou, H. Yuan, X. Xing, R. Yan, C. Li, S. Wang, T. Huang, Z. Liu, Omnigen: unified image generation, in: CVPR, 2025, pp. 13294–13304.
[100] Q. Cai, J. Chen, Y. Chen, Y. Li, F. Long, Y. Pan, Z. Qiu, Y. Zhang, F. Gao, P. Xu, et al., HiDream-I1: a high-efficient image generative foundation model with sparse diffusion transformer, arXiv preprint arXiv:2505.22705, 2025.
[101] Y. Deng, X. He, C. Mei, P. Wang, F. Tang, Fireflow: fast inversion of rectified flow for image semantic editing, arXiv preprint arXiv:2412.07517, 2024.
[102] Q. Zhang, M. Chen, A. Bukharin, P. He, Y. Cheng, W. Chen, T. Zhao, Adaptive budget allocation for parameter-efficient fine-tuning, in: ICLR, 2023, pp. 1–17, https://openreview.net/forum?id=lq62uWRJjiY.
[103] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.
[104] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, D. Cohen-Or, Prompt-to-prompt image editing with cross attention control, arXiv preprint arXiv:2208.01626, 2022.
[105] Z. Huang, T. Wu, Y. Jiang, K.C. Chan, Z. Liu, ReVersion: diffusion-based relation inversion from images, in: SIGGRAPH Asia 2024 Conference Papers, 2024, pp. 1–11.

Author biography



Jiun Tian Hoe is a Ph.D. candidate with the Centre for Advanced Robotics Technology Innovation (CARTIN) in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he is advised by Prof. Yap-Peng Tan and Prof. Xudong Jiang. He received the Bachelor of Computer Science degree from the Faculty of Computer Science and Information Technology, Universiti Malaya, Malaysia, in 2022. His research interests include computer vision, image generation, image editing, world models and generative models. He regularly serves as a reviewer for top-tier computer vision and artificial intelligence venues, including CVPR, ICCV, ECCV, NeurIPS, BMVC, and AAAI.



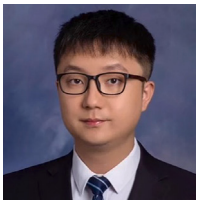
Weipeng Hu received the Ph.D. degree in Electronics and Information Technology from Sun Yat-sen University, Guangzhou, China, in 2022. From 2022 to 2026, he was a Research Fellow with the Centre for Advanced Robotics Technology Innovation (CARTIN) in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He is currently an Associate Professor with the School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, China. His research interests include computer vision, intelligent perception and understanding, and multimodal information processing. He has published over 30 papers in top-tier journals and conferences. Dr. Hu serves as an Area Chair for IEEE ICME 2026 and BMVC 2026, and regularly reviews for prestigious venues including IEEE T-PAMI, T-IP, T-IFS, T-MM, T-CSVT, and CVPR.



Wei Zhou received the M.E. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2020, and the Ph.D. degree from the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China, in 2024. His research interests include computer vision, pattern recognition, and deep learning, with a specific focus on multi-label image classification, image captioning, weakly supervised temporal action localization, and image generation and editing.



Chao Xie received the Ph.D. degree from Southeast University, Nanjing, China, in 2018. From July 2023 to July 2024, he is a Visiting Scholar with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He is currently an Associate Professor with the College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing. His current research interests include image processing and deep learning.



Ziwei Wang is currently an assistant professor in School of Electrical and Electronic Engineering, Nanyang Technological University and the director of Perception and Embodied INtElligence (PINE) Lab. Before joining NTU, he was a postdoctoral fellow in Robotics Institute, Carnegie Mellon University, with Prof. Changliu Liu. He received the Ph.D. and the B.S. degrees from the Department of Automation, Tsinghua University in 2023 and the Department of Physics, Tsinghua University in 2018 respectively. His research goal is to design foundation models (FMs) for robotic manipulation. He has published over 50 scientific papers in TPAMI, IJCV, RAL, CVPR, ICCV, ECCV, NeurIPS, CoRL, IROS and ICRA. He serves as a regular reviewer for a variety of conferences and journals.



Xudong Jiang (Fellow, IEEE) received the BE and PhD degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, and the PhD degree from Helmut Schmidt University, Hamburg, Germany. From 1998 to 2004, he was with the Institute for Infocomm Research, A*STAR, Singapore, as a lead scientist, and the head of the Biometrics Laboratory. He joined Nanyang Technological University (NTU), Singapore, as a faculty member, in 2004, where he served as the director of the Centre for Information Security from 2005 to 2011. He is currently a professor with the School of EEE, NTU and serves as the director of the Centre for Information Sciences and Systems of School of EEE, NTU. He holds 7 patents and has authored about 300 papers with more than 60 papers in the IEEE journals and more than 30 papers in top conferences such as CVPR/NeurIPS/ICCV/ECCV/ICLR/AAAI. He served as IFS TC member of the IEEE Signal Processing Society from 2015 to 2017, associate editor for IEEE Signal Processing Letters from 2014 to 2018 and associate editor for IEEE Transactions on Image Processing from 2016 to 2020. Currently, he serves as senior area editor for IEEE Transactions on Image Processing and editor-in-chief for IET Biometrics. His research interests include image processing, pattern recognition, computer vision, machine learning, and biometrics.



Yap-Peng Tan (Fellow, IEEE) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1993, and the M.A. and Ph.D. degrees in electrical engineering from Princeton University, Princeton, New Jersey, in 1995 and 1997, respectively. He is currently the Provost at VinUniversity, Hanoi, Vietnam, and a Full Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include image and video processing, computer vision, pattern recognition, and data analytics. Prof. Tan has served as an Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology, IEEE Signal Processing Letters, IEEE Transactions on Multimedia, and IEEE Access. He was also an editorial board member for the EURASIP Journal on Advances in Signal Processing and the EURASIP Journal on Image and Video Processing. His conference leadership includes serving as General Co-Chair for ICME 2010 and VCIP 2015, and Technical Program Co-Chair for ICME 2015 and ICIP 2019.



Chee Seng Chan (Senior Member, IEEE) is a Full Professor at the Faculty of Computer Science and Information Technology, Universiti Malaya, and an Affiliate Faculty at VinUniversity. His work spans computer vision, machine learning, and trustworthy artificial intelligence. He leads an active research team whose work has been published in leading international venues such as CVPR, NeurIPS etc. He is among the key architects of ILMU, Malaysia's first sovereign multilingual large language model, and has played an important role in advancing locally grounded, strategically important AI systems ready for real-world deployment. His contributions include Ryt AI, one of the earliest regulator-approved, language-centric AI systems for banking, and MalayMMLU, the first benchmark for evaluating AI proficiency in Bahasa Melayu. He has received several recognitions, including the Top Research Scientists Malaysia (TRSM) Award, the Young Scientists Network of the Academy of Sciences Malaysia (YSN-ASM), and the Hitachi Research Fellowship. He has held leadership roles in major international conferences, including IEEE VCIP 2013, ACPR 2015, IEEE MMSP 2019, and ACM MMAsia 2025 and 2026. From 2020 to 2022, he served as Undersecretary for the Division of Data Strategic and Foresight at Ministry of Science, Technology and Innovation (MOSTI).