OneHOI: Unifying Human-Object Interaction Generation and Editing

Jiun Tian Hoe[1] Weipeng Hu[1,2*] Xudong Jiang[1] Yap-Peng Tan[1,4] Chee Seng Chan[3*]
[1]Nanyang Technological University [2]Sun Yat-sen University [3]Universiti Malaya [4]VinUniversity

Code and dataset: https://jiuntian.github.io/OneHOI/

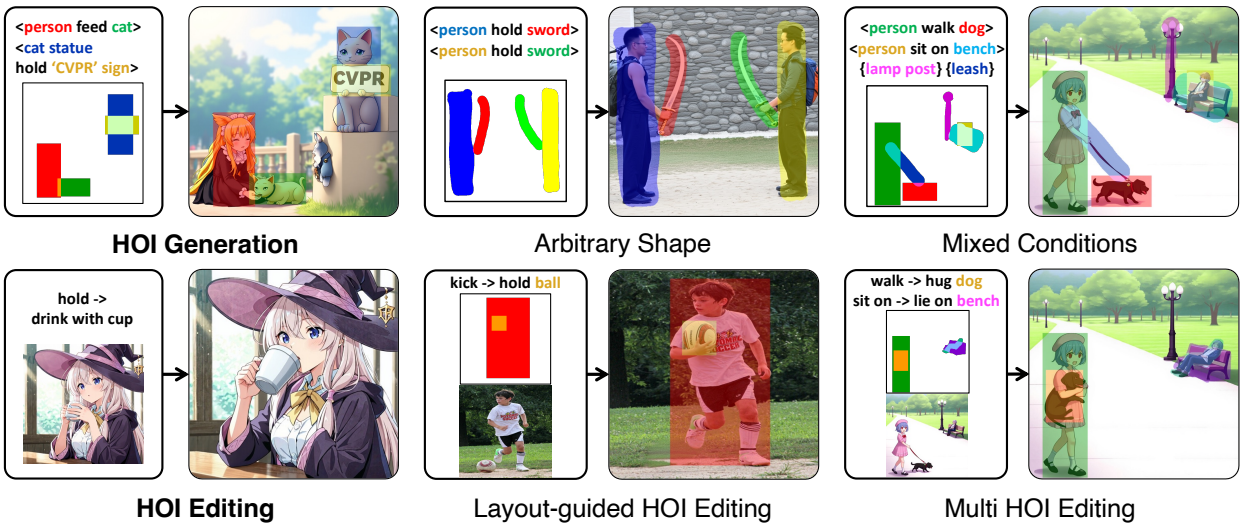


Figure 1. **OneHOI** unifies Human-Object Interaction (HOI) generation and editing in a single, versatile model. It excels at challenging HOI editing, from text-guided changes to novel layout-guided control and novel multi-HOI edits. For generation, **OneHOI** synthesises scenes from text, layouts, arbitrary shapes, or mixed conditions, offering unprecedented control over relational understanding in images.

Abstract

*Human-Object Interaction (HOI) modelling captures how humans act upon and relate to objects, typically expressed as ⟨person, action, object⟩ triplets. Existing approaches split into two disjoint families: HOI generation synthesises scenes from structured triplets and layout, but fails to integrate mixed conditions like HOI and object-only entities; and HOI editing modifies interactions via text, yet struggles to decouple pose from physical contact and scale to multiple interactions. We introduce **OneHOI**, a unified diffusion transformer framework that consolidates HOI generation and editing into a single conditional denoising process driven by shared structured interaction representations. At its core, the Relational Diffusion Transformer (R-DiT) models verb-mediated relations through role- and instance-aware HOI tokens, layout-based spatial Action Grounding, a Structured HOI Attention to enforce interaction topology, and HOI RoPE to disentangle multi-HOI scenes. Trained jointly with modality dropout on our HOI-Edit-44K, along with HOI and object-centric datasets, **OneHOI** supports layout-guided, layout-free, arbitrary-mask, and mixed-condition control, achieving state-of-the-art results across both HOI generation and editing.*

*Corresponding authors: Weipeng Hu (huwp7@mail.sysu.edu.cn) and Chee Seng Chan (cs.chan@um.edu.my)

1. Introduction

Human-Object Interaction (HOI) lies at the forefront of visual understanding, focusing not just on what appears in an image but also on how entities relate. It represents the world through structured triplets ⟨person, action, object⟩, capturing the grammar of interaction. Mastering HOI is crucial for next-generation AI, from building dynamic AR/VR worlds to enabling content creation that understands why and how things connect, not merely what they are.

Existing studies follow two main directions. *Recognition and detection* approaches [3, 6, 26] aim to identify and localize HOI, improving perceptual understanding but offering no generative capability. *Generative methods* [4, 13, 14, 43] in contrast, have evolved into two disjoint families: **HOI generation**, which synthesises scenes from triplets conditioned on spatial layouts for controllability, but struggles with flexible control, such as *mixing HOI triplets*

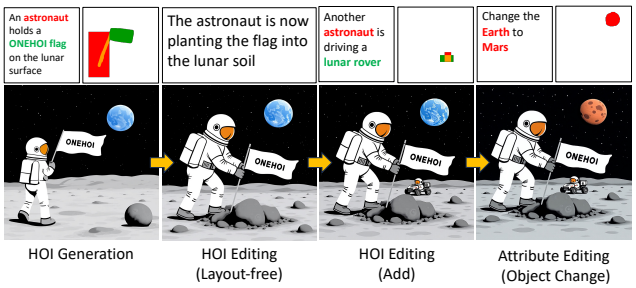


Figure 2. **Unified HOI generation and editing. OneHOI** enables a single-model multi-step workflow. It begins with (i) **Mixed-Condition Generation**, synthesising a complex scene from layout-guided HOIs with arbitrary shape. Then, it performs (ii) **Layout-free HOI Editing**, (*e.g.*, change him to plant the flag), followed by (iii) **Layout-guided HOI Editing** (*e.g.*, add another astronaut and driving a rover) and (iv) **Attribute Editing** (*e.g.*, change to Mars). More examples in Fig. 14 of the Appendix.

with object-only entities or accepting *arbitrary shape* layouts; and **HOI editing**, which modifies images via text, cannot reliably decouple and recompose pose and physical contact. Besides, it fails to scale beyond a single interaction, lacks fine spatial control, and relies on implicit priors rather than explicit structural modelling.

This paper asks a simple but fundamental question: ***Can HOI generation and editing be unified within a single framework?*** We posit that joint training creates a substantial synergy, as the broad interaction semantics (*e.g.*, poses, contact points) learned during generation can provide the deep structural HOI knowledge that editing-only models lack, enabling more plausible and physically-aware edits.

Achieving this unification requires a high-fidelity backbone with a flexible architecture for multi-modal conditioning. Diffusion Transformers (DiTs) [30] are a promising candidate. They combine diffusion's fidelity with transformers' global reasoning to produce high-quality images [9, 21, 38] and enable fine-grained spatial control [45]. Yet, they have a critical flaw: *DiTs treat scenes as collections of independent objects and lack explicit interaction modelling, yielding visually detailed but relationally shallow results.*

To address this, we introduce **OneHOI**, a unified framework for HOI generation and editing. Our key insight is that both tasks are two views of a single conditional denoising process. Besides layouts and captions, our model also conditions on structured interaction representations, reframing diffusion from arranging pixels to realising relationships.

At the core of **OneHOI** lies a new Relational DiT (R-DiT) with three tightly coupled modules: (i) *HOI Encoder* to inject role- and instance-aware cues into HOI token; (ii) *Structured HOI Attention* to enforce a verb-mediated topology among HOI tokens and (iii) *HOI RoPE* to assign distinct positional identities to disentangle interactions in multi-HOI scenes. Together, these form a unified grammar that enables reasoning over interactions, not just regions.

Trained jointly for generation and editing on our new HOI-Edit-44K dataset with modality-dropout, supplemented by established HOI and object-level datasets, **OneHOI** is a unified pipeline that supports layout-guided, layout-free, arbitrary-mask, and mixed-condition controls, handling single and multiple interactions, see Figs. 1 and 2.

Our main contributions are:

- **OneHOI**, a unified DiT-based framework for HOI generation and editing, scaling to multi-HOI scenes and, for the *first time*, enabling multi-HOI editing.
- A novel R-DiT that embeds explicit interaction representations via three modules (*i.e.* HOI Encoder, Structured HOI Attention, and HOI RoPE), enabling precise yet flexible control under diverse conditions, including layout-guided, layout-free, arbitrary masks, and mixed inputs.
- A new large-scale paired dataset, HOI-Edit-44K, addressing the scarcity of paired data, with 44K identity-preserving examples, for training of robust HOI editing.
- State-of-the-art performance across benchmarks for controllable HOI generation, layout-free editing, and novel layout-guided single- and multi-HOI editing tasks.

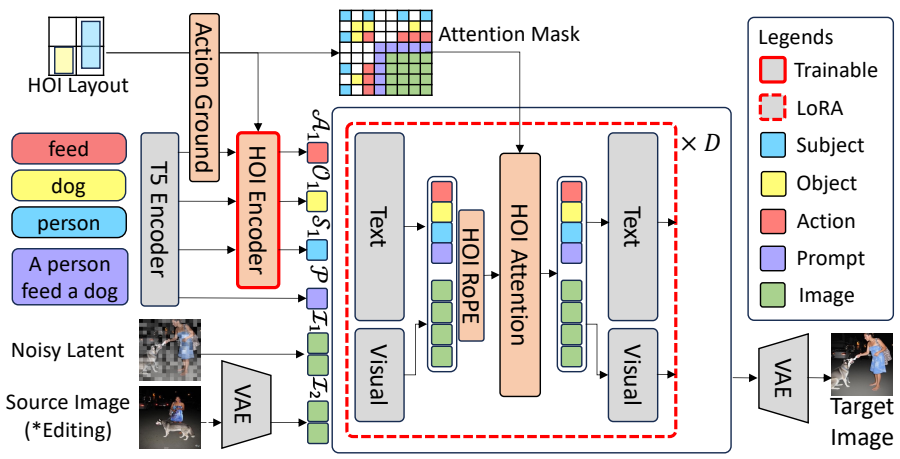
2. Related Works

Controllable Generation and Human-Object Interaction. Research on fine-grained control [23] and spatial conditioning (e.g., GLIGEN [22], MIGC [47] and EliGen [45]) has enabled object placement via layouts or attention manipulation. However, they focus on individual entities, specifying *where* objects are, but not *how they relate*. Generative HOI research addresses this gap, diverging into:

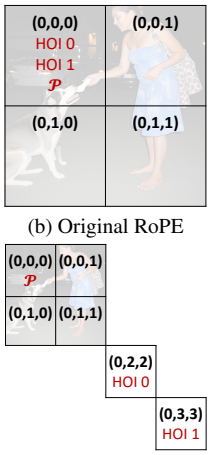
- **Layout-Conditioned Generation.** Methods like InteractDiffusion [13] synthesise images from triplets conditioned on spatial layouts for controllability, but struggle with flexible control (*e.g.*, *mixing HOI triplets* with object-only entities or accepting *arbitrary shape* layouts) and fail when layout guidance is partial or absent.
- **Text-Guided Editing.** Methods like HOIEdit [43] and InteractEdit [14] modify interactions in existing images. They cannot reliably decouple and recompose the pose and physical contact, fail to scale beyond a single interaction, lack precise spatial control, and rely on implicit model priors rather than explicit interaction modelling.

This fragmented development leaves a clear gap: no unified framework bridges these modalities and the multi-HOI editing is largely unaddressed. **OneHOI addresses these limitations directly**, introducing the first framework to unify generation and editing, enabling precise yet flexible control under diverse conditions (*e.g.*, layout-guided, layout-free, arbitrary masks, and mixed inputs) within one model.

Diffusion Transformers (DiTs) for Image Synthesis. The landscape of image synthesis has been reshaped by diffusion models [12, 35], which have rapidly surpassed GANs in generating high-fidelity images. Latent Diffusion Models



(a) An overview of **OneHOI** pipeline.



(b) Original RoPE

(c) HOI RoPE

Figure 3. **(a) OneHOI** unifies HOI editing and generation tasks on a DiT backbone. The pipeline features an HOI Encoder to inject role and instance cues, and Structured HOI Attention to enforce verb-mediated topology and spatial grounding. **(b, c)** To separate instances, in contrast to the Original RoPE (b), HOI RoPE (c) provides unique positional indices for each interaction.

[33] democratised this by operating in a compressed latent space, significantly reducing computational costs. While early models used U-Nets [34], DiTs [30] marked a pivotal shift. Replacing convolutions with a pure transformer architecture yielded superior scaling properties, establishing DiTs as the new standard. State-of-the-art systems like Flux.1 [21] and Qwen-Image [38] leverage Multi-Modal DiT (MM-DiT) [9] variants with flow-matching objectives [24], achieving unprecedented quality and controllability, yet they lack explicit interaction modelling.

3. Methodology

Figure 3a overviews our unified pipeline for HOI generation and editing. Given a global text prompt $\mathcal{P}$ and either a set of structured interaction $\{\langle s, o, a\rangle_n\}_{n=1}^N$ or independent objects $\{\langle o\rangle_n\}_{n=1}^N$ with optional layout $\mathcal{B} = \{b_n^s, b_n^o\}$ or $\mathcal{B} = \{b_n^o\}$, our pipeline produces an image that realises all specified targets. We denote the sets of T5 [31]-encoded tokens corresponding to these triplets as $\mathcal{H} = \bigcup_{n=1}^N \{\mathcal{S}_n, \mathcal{A}_n, \mathcal{O}_n\}$, where $\mathcal{S}_n, \mathcal{O}_n, \mathcal{A}_n$ represent subject, object, and action tokens, respectively for instance n. For generation, we sample noise $\mathcal{I}_1$ in the latent space and run the conditional denoiser. For editing, we encode the source image into latents $\mathcal{I}_2$, concatenate them with the noise $\mathcal{I}_1$, and run the *same* denoiser conditioned on the new interaction targets.

Our core idea is the introduction of Relational DiT (R-DiT), a modified backbone that *explicitly models interaction structure*. We build the R-DiT by introducing four key components to a standard layout-conditioned DiT baseline, Eligen [45], as validated in our ablation (Sec. 4.7). These components inject increasingly sophisticated relational understanding: (i) **Action Grounding**, which introduces action-specific semantic and spatial cues; (ii) **HOI Encoder**, adding fine-grained role and instance identity; (iii) **Structured HOI Attention**, enforcing a verb-mediated attention topology and layout constraints; and (iv) **HOI RoPE**, ensuring interaction instances separation in complex scenes. More details in Appendix A.

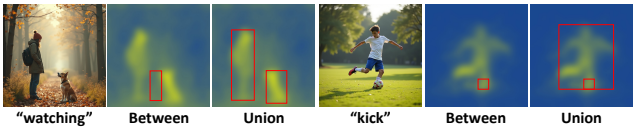


Figure 4. Action-token→image attention heatmaps from the baseline. The "Between" region proposed in InteractDiffusion [13] misses where the action actually attends, while our "Union" region (subject ∪ object) better matches the attention footprint.

3.1. Action Grounding

Standard layout-conditioned models only ground objects. To model interactions, however, the model must also have basic awareness of the *action* itself, both semantically and spatially. We introduce *Action Grounding* (AG) to provide this foundational capability. It builds upon a baseline that grounds subject $\mathcal{S}_n$ and object $\mathcal{O}_n$ tokens to regions R_n^s and R_n^o by introducing two action-specific cues: **(i) Semantic Action Token** $\mathcal{A}_n$ (T5 [31]-encoded) for each action label (*e.g.* "feed") in the HOI triplet and **(ii) Spatial Action Region** R_n^a associated with this action.

Previous work [13] defines the action region with a "between" operator, which uses the intersection of the subject and object boxes when they overlap, else a rectangle spanning them when they disjoint. While adequate as a conditioning cue, this band often fails to match *where* the action token actually attends (too narrow or misplaced; see Fig. 4).

We define it instead as the **union** of the subject and object regions. By rasterising the subject and object shapes/boxes

b_n^s, b_n^o, we form regions R_n^s and R_n^o, and set $R_n^a = R_n^s \cup R_n^o$. This choice (i) aligns better with the natural attention patterns of DiT, (ii) is robust for both overlapping and disjoint pairs, and (iii) provides a stable target for grounding the action (via Sec. 3.3). This establishes the foundational understanding of interaction that is missing in object-only models, upon which our subsequent modules are built.

3.2. HOI Encoder

Models risk *role confusion* or *blending wrong interactions* in multi-HOI scenes. For example, given ⟨person1, chase, dog⟩ and ⟨person2, hold, cat⟩, a model might incorrectly render 'person1' *holding* the 'cat' (**blending wrong interactions**) or a *dog chasing* 'person1' (**role confusion**). Hence, simply providing $\mathcal{S}_n, \mathcal{O}_n, \mathcal{A}_n$ tokens is insufficient. The model must explicitly know *which token plays which role* (subject/object/action) and *which interaction instance* it belongs to. HOI Encoder tackles this by injecting compact, explicit identity cues into the HOI token streams $\mathcal{H}$.

Formulation. Let d as T5 output dimension (d=4096). For an interaction instance n and role $r \in \{s, o, a\}$, let $x_n^r \in \mathbb{R}^d$ be the T5-embedding. We build three side signals:

$$e_{\text{role}}(r) \in \mathbb{R}^{64}, \quad e_{\text{inst}}(n) \in \mathbb{R}^{64}, \quad e_{\text{box}}(b_n^r) \in \mathbb{R}^{256},$$

where $e_{\text{role}}(r)$ is a learnable role embeddings, $e_{\text{inst}}(n)$ is a fixed sinusoidal embedding of the instance index, and $e_{\text{box}}(b_n^r)$ is Fourier embedding [27] of the role's box.

We then normalize the HOI token h_n^r with Layer Normalization, concatenate it with the side signals and project the result with a small MLP, and apply a gated residual:

$$\tilde{h}_n^r = \text{MLP}([\text{LN}(h_n^r); e_{\text{box}}(b_n^r); e_{\text{role}}(r); e_{\text{inst}}(n)]), \quad (1)$$

$$\tilde{h}_n^r = h_n^r + \tanh(\lambda) \cdot \tilde{h}_n^r, \quad (2)$$

where $\lambda \in \mathbb{R}$ is a learnable gate that smoothly ramps in the conditioning to stabilise training. The augmented tokens $\tilde{h}_n^r$ are then fed into the DiT backbone. This provides the fine-grained identity information necessary for multi-HOIs relational modelling.

3.3. Structured HOI Attention

Standard layout conditioning often treats subjects and objects as *independent entities*. This means it can place them correctly but fails to capture the interaction structure, as it ignores the specific semantic and geometric relationship dictated by the *action*. This independence leads to plausible but incorrect outputs, such as failing to render the 'holding' interaction in Fig. 10-(2) or generating other awkward poses. We introduce Structured HOI Attention to explicitly embed this relational structure via a *verb-mediated* attention topology. It governs attention patterns via masking, controlling both how HOI tokens $\mathcal{H}$ interact amongst themselves and how they ground to the image $\mathcal{I}$.

HOI↔HOI Topology. Our key insight is that action is central to defining the interaction structure. For each instance

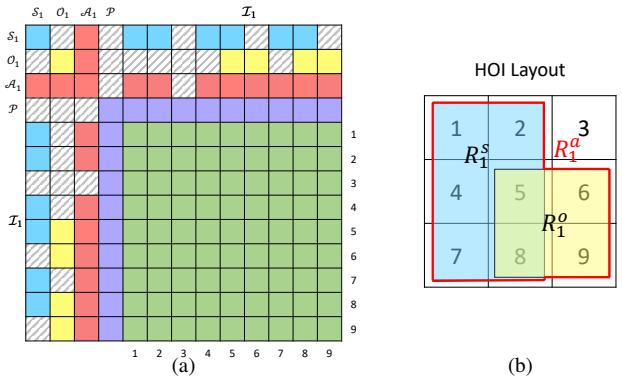


Figure 5. (a) HOI attention mask. Colours match Fig. 3a legend, grey hatched indicates blocked attention. Direct $\mathcal{S}_n \leftrightarrow \mathcal{O}_n$ is blocked to enforce verb-mediated topology. $\mathcal{S}_n, \mathcal{O}_n, \mathcal{A}_n$ attend to image $\mathcal{I}_1$ only within R_n^s, R_n^o, R_n^a, respectively, as shown in (b).

n, we prevent the direct links between subject↔object and enforce a verb-mediated pathway (cf. top-left of Fig. 5):

$$\mathcal{S}_n \leftrightarrow \mathcal{A}_n, \qquad \mathcal{O}_n \leftrightarrow \mathcal{A}_n, \qquad \textit{block } \mathcal{S}_n \leftrightarrow \mathcal{O}_n.$$

All cross-instance HOI links ($n \neq m$) are also disabled. This forces relational information to flow through the action tokens $\mathcal{A}_n$, directly reflecting the interaction's structure.

HOI↔Image Grounding. When layout is provided, we constrain HOI→image attention between HOI query $q \in \{\mathcal{S}_n, \mathcal{A}_n, \mathcal{O}_n\}$ and image key $k \in \mathcal{I}$ as:

$$M_{\mathcal{H}\mathcal{I}}(q,k) = \begin{cases} 0, & q \in \mathcal{S}_n \text{ and } k \in R_n^s, \\ 0, & q \in \mathcal{O}_n \text{ and } k \in R_n^o, \\ 0, & q \in \mathcal{A}_n \text{ and } k \in R_n^a, \\ -\infty, & \text{otherwise.} \end{cases} \quad (3)$$

This rule applies symmetrically for image→HOI attention. When layout is absent, these constraints are removed (all connections allowed). This component compels the model to learn the semantic and spatial structure of the interaction.

Final Attention. The attention mask $\mathcal{M}$ (Fig. 5) aggregates (i) the **HOI↔HOI topology**, (ii) the **HOI↔image grounding** constraints $M_{\mathcal{H}\mathcal{I}}$, and (iii) the standard connections for prompt↔image and image↔image. The prompt↔HOI tokens are blocked. The final attention is:

$$\text{Attn}(Q, K, V, \mathcal{M}) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}} + \mathcal{M}\right) V, \quad (4)$$

with $\mathcal{M}_{qk} = 0$ for allowed pairs and a large negative value (implementing $-\infty$) otherwise.

3.4. HOI RoPE (HRoPE)

Processing multi-HOIs simultaneously risks "cross-talk", where feature from one instance leaks and influences another, causing blended interactions or attributes swap. For instance, given ⟨person1, chase, dog⟩ and ⟨person2, hold, cat⟩, cross-talk might cause the model to generate "person1

holding the cat", incorrectly blending the two instances. HOI RoPE is a specialized positional indexing scheme to separate interaction instances. It is applied to the query Q and key K for all HOI tokens $\mathcal{H}$ in the attention (Eq. (4)). The image stream uses 3D RoPE [36] over a spatial grid of size $H \times W$ following [21]. We assign all HOI tokens $\mathcal{H}$ belonging to the same instance n a single, distinct positional index from the image grid and other instances:

$$z_{\text{HOI}}(n) = (0,\ T{+}n, T{+}n), \quad \text{where} \quad T = \max(H, W).$$

This assigns each interaction a unique "slot" in the RoPE space (cf. Fig. 3c). Applied across all layers, HRoPE reduces inter-instance interference in multi-HOI scenes.

4. Experiments

We implement **OneHOI** by adapting the MM-DiT backbone from Flux.1 Kontext [21]. We train using LoRA [15] for 10K steps with a batch size of 16 using the AdamW [17] optimizer (8-bit). More details are provided in Appendix A, see Appendix C.2 for human preference study.

4.1. Unified Training Strategy

To enable a single model for both generation and editing under diverse conditions, we employ a joint training strategy with modality dropout. Batches alternate between **generation** and **editing** and we optimize with the standard diffusion flow-matching objective [24]. During training, we randomly drop input modalities: layout (bounding boxes b_n^r) with probability $p_{\text{layout}} = 0.25$, HOI labels ($\langle s, o, a \rangle_n$ replaced by object-only) with $p_{\text{hoi}} = 0.25$, and the global text prompt $\mathcal{P}$ with $p_{\text{txt}} = 0.30$, ensuring at least one modality remains. The attention masking (Sec. 3.3) is applied consistently, defaulting to unconstrained attention for dropped layouts. This ensures the model operates robustly across various tasks and input combinations.

4.2. Datasets

HOI-Edit-44K (ours). To address the lack of paired data for HOI editing, we constructed a large-scale dataset, HOI-Edit-44K, which we will release publicly. We collect source images with verified HOIs from two streams: (i) **Flux.1 generations** that realise a verified source interaction, and (ii) **HICO-DET images**. For each source image, we synthesise potential single-HOI edits using Flux.1 Kontext [21] and InteractEdit [14]. A (source, edited) image pair is retained only upon passing two rigorous automated checks:

- **HOI correctness.** We run PViC [44] HOI detector on the edited image and require the predicted HOI to match the target HOI. The detected layout are recorded for the pair.
- **Identity preservation.** We extract DINOv2 features [29] from subject and object crops in both source and edited images and keep the pair only if both cosine similarities exceed a threshold of 0.75.

This stringent filtering process discarded approximately 90% of initial candidates, primarily due to incorrect interactions or identity drift. The final dataset comprises 44K high-quality HOI editing pairs, each including the source images, target interaction triplet, edited image, and corresponding layout. This provides diverse, identity-preserving interaction edits at scale, crucial for training our unified model. See Appendix B.3 for more details and generalization.

SA-1B [18]. We sample 35K images and derive layouts from object masks [23], providing *object-only layout* supervision (no HOI) that strengthens spatial layout control.

HICO-DET [5]. We use 37K training images to learn HOI generation priors. The test set is used only for evaluation.

4.3. Metrics

Image Quality. We report standard human-preference aligned perceptual metrics: **PickScore**[19], **HPSv2** [40] (Human Preference Score), and **ImageReward**[42]. Higher scores indicate better quality and prompt alignment.

HOI Editing. As to [14], we report **HOI Editability** (success of the target verb–object being realised, as detected) and **Editability–Identity** (a composite that balances HOI success with ID preservation). See Appen. C.1 for details.

Spatial Score. For layout-guided tasks, we run PViC [44] to detect subject and object instances for each target triplet. We compute the mean IoU between the target boxes (b^s, b^o) and the best-matching detected boxes $(\hat{b}^s, \hat{b}^o)$, defined as $\text{mIoU} = \frac{1}{2}\big(\text{IoU}(b^s, \hat{b}^s) + \text{IoU}(b^o, \hat{b}^o)\big)$. Results are averaged over all targets, higher means better spatial alignment.

HOI Accuracy. Using PViC, a success is recorded when the target HOI is detected within their specified regions. We report the mean success rate across targets (higher is better).

4.4. Tasks and Evaluations

We evaluate three HOI tasks differ by available controls:

Layout-free HOI editing. Modifying interactions in an image using only HOI triplets (no layout), while preserving identity and image quality. We generate 1000 samples for 100 target edits in IEBench [14] and report Editability–Identity, HOI Editability, and image quality metrics (PickScore, HPS, ImageReward).

Layout-guided HOI editing. Modifying interactions in an image using HOI triplets and target layouts. With layout guidance, it enable editing multiple HOIs at once, which was challenging due to limited expressibility in natural language. For single-HOI edits, we use IEBench with synthesised target layouts, detailed in Appendix B.1. For Multi-HOI edits, we *propose a new MultiHOIEdit benchmark* (detailed in Appendix B.2), comprises of 200 target edits spanning 2–3 interactions per image, where we generates total 1000 samples for evaluation. In addition to the layout-free metrics, we also report Spatial Score.

HOI generation. Synthesising images from HOI triplets

Table 1. Quantitative comparison for **layout-free HOI editing** on IEBench benchmark. Our method significantly outperforms others across all metrics for editing and image quality. Best results are in **bold**, second best are underlined. Final row shows the closed-source baseline.

Method	HOI Editing		Image Quality		
	Editability-Identity	HOI Editability	PickScore	HPS	ImageReward
Null-Text Inversion [10, 28]	0.443	0.390	20.81	0.2483	−0.3329
MasaCtrl [2]	0.371	0.260	20.14	0.2212	−0.7136
HOIEdit [43]	0.349	0.240	19.51	0.2129	−1.0289
InstructPix2Pix [1]	0.380	0.269	20.28	0.2178	−0.7717
TurboEdit [8]	0.434	0.326	20.36	0.2437	−0.3821
EditFriendlyDDPM [16]	0.438	0.320	20.48	0.2470	−0.3875
OmniGen [41]	0.354	0.231	19.74	0.2120	−1.0055
FireFlow [7]	0.451	0.350	20.76	0.2530	−0.4385
Flux.1 Kontext [21]	0.471	0.328	20.45	0.2427	−0.5137
OmniGen2 [39]	0.496	0.437	20.90	0.2595	−0.0869
Qwen Image Edit [38]	0.580	0.460	20.81	0.2585	0.0748
InteractEdit [14]	0.573	0.514	21.08	0.2640	0.1630
Ours	**0.638**	**0.596**	**21.26**	**0.2805**	**0.4713**
Improvements	10.0%	16.0%	0.85%	6.25%	189%
Nano Banana	0.623	0.530	20.97	0.2544	0.1810

Table 2. Quantitative results for our novel **layout-guided HOI editing** tasks. We report strong performance for both single- and multi-HOI editing, establishing the first baseline for these new capabilities.

Task	Method	Layout-guided HOI Editing			Image Quality		
		Editability-Identity	HOI Editability	Spatial	PickScore	HPS	ImageReward
Single HOI Editing	InteractEdit + InteractDiffusion	0.559	0.520	0.749	20.53	0.2418	−0.3072
	Ours	**0.638**	**0.570**	**0.822**	**21.04**	**0.2678**	**0.2897**
Multi HOI Editing	Ours*	0.435	0.329	0.675	21.22	0.2742	0.1954

* There is no other baseline that performs layout-guided multi-HOI editing task, thus we report only ours.

Table 3. Quantitative comparison for **HOI generation** task. Our method outperforms leading layout-conditioned and HOI-aware models on both controllability and image quality metrics.

Method	Controllability		Image Quality		
	Spatial	HOI	PickScore	HPS	ImageReward
GLIGEN [22]	0.5150	0.3344	20.46	0.2322	−0.4103
InstanceDiffusion [37]	0.5228	0.3476	20.06	0.2312	−0.2532
MIGC++[46, 47]	0.5331	0.3616	20.16	0.2208	−0.6492
Eligen [45]	0.4371	0.3061	21.28	0.2496	0.3921
InteractDiffusion [13]	0.5768	0.4505	20.37	0.2283	−0.3194
Ours	**0.6104**	**0.4528**	**21.41**	**0.2617**	**0.5224**
Improvements	5.8%	0.5%	0.6%	4.8%	33.2%

and layouts. We evaluate on 2000 HICO-DET test targets and report HOI accuracy, Spatial score, and image quality.

4.5. Quantitative Results

Layout-free HOI editing. Table 1 compares our method with recent editing baselines. We achieve the best Editability–Identity (0.638) and HOI Editability (0.596), improving over the strongest priors by +10.0% and +16.0%, respectively, while also attaining the best HPS, ImageReward and PickScore. These results indicate that, even without layout input, our unified formulation reliably edits the interaction while maintaining subject identity intact.

Layout-guided HOI editing. Table 2 reports single- and multi-HOI edits with layout guidance. For single-HOI, we establish a baseline by adapting InteractEdit[14] and InteractDiffusion [13] (see Appendix A.3). Our method achieves a high Spatial score (0.822), strong HOI Editability (0.570), and good perceptual quality. For much harder multi-HOI (2–3 HOIs across 1-3 persons), Spatial remains strong (0.675) and quality scores are maintained.

HOI generation. Table 3 reports controllability and perceptual quality. Our method slightly surpasses [13] on Spatial and HOI accuracy, while also achieving the best perceptual scores, PickScore 21.41 (+0.7%), HPS 0.2617 (+4.8%) and ImageReward 0.5524 (+33.2%) over the strongest prior. Thus, unifying editing and generation does not compromise HOI generation; instead, it improves it.

4.6. Qualitative Results

Fig. 6 compares **layout-free HOI editing**. HOIEdit [43] often corrupts the image. For *hold→ride skateboard*, Qwen leaves the pose essentially unchanged and [14] drifts in identity; others have an incorrect riding stance. Contrary, **OneHOI** renders the intended interaction while preserv-

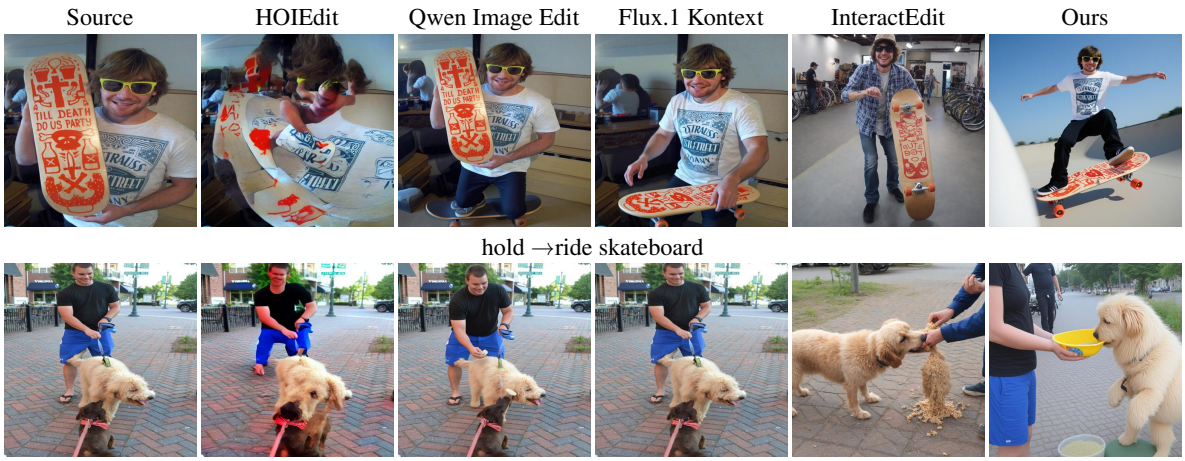


Figure 6. Qualitative comparison for layout-free HOI editing. Our method successfully renders the new interaction while preserving identity. In contrast, baseline methods often produce artifacts, fail to change the pose, or lose the subject's identity.

Figure 7. Qualitative comparison for HOI generation. While object-level methods correctly place entities, they fail to synthesise specified interactions. Ours renders semantically and geometrically consistent interactions, demonstrating a deeper relational understanding.

ing identity. This stems from two separate factors: (i) HOI semantics learned during generation (contact patterns, verb–object geometry) transfer to editing, and (ii) structured HOI attention steers the edit to the correct roles/regions. Baselines without such HOI knowledge tend to keep poses unchanged or misrender contact.

Fig. 7 compares **HOI generation**. *Object-level methods* (GLIGEN, MIGC, InstanceDiff, Eligen) correctly place entities but rarely realise the relations: not texting on the phone. For *HOI-level*, [13] improves relation plausibility but often produces less convincing, semantically off interactions. Our model, **OneHOI** yields superior semantic faithfulness, *e.g.*, hands grasp the phone for 'holding/reading/texting,'. We attribute these gains to: (i) **HOI tokens** that encode the interaction semantics, (ii) **structured HOI attention** that constrains HOI tokens to their regions while models the relation, and (iii) **HOI RoPE** that separates instances to avoid mix-ups. This yields spatially compliant and semantically faithful multi-HOI scenes.

Fig. 8 shows **layout-guided HOI edits**. For single-HOI scene: the edits are confined to the layout. The ball is firmly grasped, and the person shifts into a riding pose on the skateboard, while their identity and background remain intact. For multi-HOI scene, natural language alone is too ambiguous to specify multiple edits; layout resolves this. Our model simultaneously executes *drink with→carry bottle* and *sit on→lie on bench*, updating each person only within their regions. One holding the bottle and the other reclining on the bench, without spillover or mix-ups. This stems from joint training with multi-HOI generation, which teaches to compose and disentangle interactions. Combined with HOI attention and HOI RoPE, this enables reliable multi-HOI edits even without multi-HOI edit training pairs.

Figure 9 showcases **arbitrary-shape masks** and **mixed-modality control**. Irregular masks (strokes/polygons) provide fine-grained shape control for subject/object regions. We combine layout-guided HOIs and object-only entities, *e.g.*, adding background props with object-only masks

Figure 8. Layout-guided editing examples. Our model supports single-HOI (top) and multi-HOI edits (bottom), limiting changes to target layouts while preserving scene consistency.

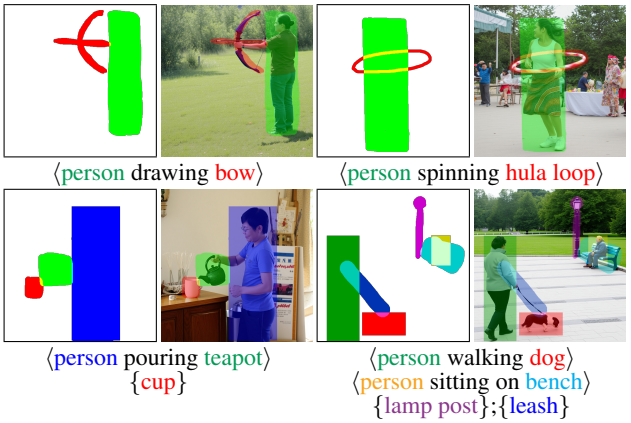


Figure 9. Versatile control in HOI generation. Our model supports conditioning on both arbitrary-shape masks (top) and a mix of HOI and object-only inputs within a single scene (bottom), demonstrating its compositional capabilities.

while generating foreground interactions. These behaviours stem from modality-dropout training and our layout-aware HOI attention. Overall, the unified interface supports flexible modality combinations in a single generation.

4.7. Ablation Studies

We conduct a comprehensive ablation study to validate the contribution of each component, summarized in Tab. 4 and visualised in Fig. 10. We perform an additive analysis, starting from a strong baseline (BL), which is the Eligen [45].

Introducing **Action Grounding (AG)** establishes a foundational understanding of interactions that the object-level model lacks. This is evident in the large gains across both generation and editing tasks. Layering on the **HOI Encoder (Enc)** further improves performance, particularly boosting the perceptual quality (IR) by providing the model with explicit role and instance cues. The subsequent addition of **Structured HOI Attention (Attn)** yields another major improvement in correctness metrics (HOI Acc. and EI), confirming its critical role in enforcing the relational structure of the interaction and adhering to layouts. Finally, incorporating **HOI RoPE (HRoPE)** provides the last refinement step by helping to disentangle instance identities, significantly enhancing perceptual quality (IR).

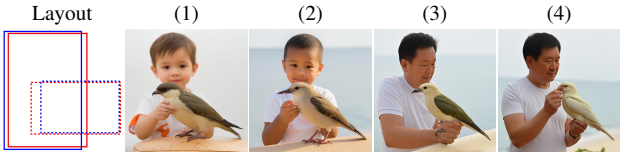


Figure 10. Progressively adding components improves the interaction's plausibility, only the full model (4) successfully rendering the complex, two-handed action of both "holding" and "petting."

Table 4. Ablation study on core components. AG: Action Grounding, Enc: HOI Encoder, Attn: HOI Attention, HRoPE: HOI RoPE, EI: Editability-Identity, IR: ImageReward.

	Components				HOI Generation		Multi-HOI Edit	
	AG	Enc	Attn	HRoPE	HOI Acc.	IR	EI	IR
BL					0.3061	0.3921	-	-
(1)	✓				0.4138	0.3156	0.423	0.1118
(2)	✓	✓			0.4254	0.4602	0.422	0.1306
(3)	✓	✓	✓		0.4504	0.4861	0.433	0.1944
(4)	✓	✓	✓	✓	**0.4528**	**0.5224**	**0.435**	**0.2046**

This progressive improvement is visualised in Figure 10 on the multi-action prompt *"A person is holding and petting bird."* (i) With only Action Grounding (AG), the model renders only a simple *'pet'* action. (ii) Adding HOI Encoder provides explicit role cues, yielding a more plausible *'petting'* pose. (iii) Introducing HOI Attention enables the *'holding'* pose but *'petting'* remains entangled with the *'holding'* gesture. (iv) Adding HRoPE separates the two action concepts and correctly depicts both *'hold'* and *'pet'*. This confirms all components are complementary in **OneHOI** for a deep relational understanding.

Appendix E shows our **unified** model outperforms **task-specific** ones via a "synergy effect", where generative priors enhance editing robustness and vice-versa.

5. Conclusion

We introduced **OneHOI**, a single DiT-based framework that unifies Human-Object Interaction (HOI) generation and editing by **explicitly modelling interaction structure**. This is realised through three core components: a dedicated **HOI Encoder** providing fine-grained role and instance identity, **Structured HOI Attention** enforcing a verb-mediated relational topology constrained by layout, and **HOI RoPE** ensuring clear instance separation. Our approach bridges the gap between layout-guided generation and layout-free editing, supports flexible control, and enables, for the first time, the challenging **multi-HOI editing** task. **OneHOI** achieves state-of-the-art controllability and perceptual quality, delivering physically plausible interactions across both editing and generation benchmarks. By effectively integrating relational structure into DiTs, our work pushes generative models beyond simple entity placement toward synthesising semantically coherent HOI scenes.

Acknowledgement

This research is supported in part by the National Research Foundation, Singapore, under the NRF Medium Sized Centre Scheme (CARTIN). Any opinions, findings and conclusions expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore. This research is also supported in part by the ASEAN-China Cooperation Fund (ACCF) under project "Deep Ensemble Under Non-Ideal Conditions and Its Typical Applications in Computer Vision."

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 6

[2] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, pages 22560–22570, 2023. 6

[3] Yichao Cao, Qingfei Tang, Xiu Su, Song Chen, Shan You, Xiaobo Lu, and Chang Xu. Detecting any human-object interaction relationship: Universal hoi detector with spatial prompt learning on foundation models. *Advances in Neural Information Processing Systems*, 36:739–751, 2023. 1

[4] SeungJu Cha, Kwanyoung Lee, Ye-Chan Kim, Hyunwoo Oh, and Dong-Jin Kim. Verbdiff: Text-only diffusion models with enhanced interaction awareness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8041–8050, 2025. 1

[5] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 5

[6] Junwen Chen and Keiji Yanai. Qahoi: Query-based anchors for human-object interaction detection. In *2023 18th International Conference on Machine Vision and Applications (MVA)*, pages 1–5. IEEE, 2023. 1

[7] Yingying Deng, Xiangyu He, Changwang Mei, Peisong Wang, and Fan Tang. Fireflow: Fast inversion of rectified flow for image semantic editing. In *ICML*, 2025. 6

[8] Gilad Deutch, Rinon Gal, Daniel Garibi, Or Patashnik, and Daniel Cohen-Or. Turboedit: Text-based image editing using few-step diffusion models. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 6

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 2, 3

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2023. 6

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2

[13] Jiun Tian Hoe, Xudong Jiang, Chee Seng Chan, Yap-Peng Tan, and Weipeng Hu. Interactdiffusion: Interaction control in text-to-image diffusion models. In *CVPR*, pages 6180–6189, 2024. 1, 2, 3, 6, 7

[14] Jiun Tian Hoe, Weipeng Hu, Wei Zhou, Chao Xie, Ziwei Wang, Chee Seng Chan, Xudong Jiang, and Yap-Peng Tan. Interactedit: Zero-shot editing of human-object interactions in images. *arXiv preprint arXiv:2503.09130*, 2025. 1, 2, 5, 6

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 5, 1

[16] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly DDPM noise space: Inversion and manipulations. In *CVPR*, pages 12469–12478, 2024. 6

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 5, 2

[19] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023. 5

[20] Black Forest Labs. Flux, 2024. GitHub repository. 1

[21] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 2, 3, 5, 6, 1

[22] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, pages 22511–22521, 2023. 2, 6

[23] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. In *arXiv preprint arXiv:2305.13655*, 2023. 2, 5

[24] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 5

[25] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ECCV*, 2024. 2

[26] Jinguo Luo, Weihong Ren, Weibo Jiang, Xi'ai Chen, Qiang Wang, Zhi Han, and Honghai Liu. Discovering syntactic interaction clues for human-object interaction detection. In *CVPR*, pages 28212–28222, 2024. 1

[27] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65:99–106, 2022. 4

[28] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, pages 6038–6047, 2023. 6

[29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5, 2

[30] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. 2, 3

[31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 3

[32] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, USA, 2nd edition, 1979. 2

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[36] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 5

[37] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6232–6242, 2024. 6

[38] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 2, 3, 6

[39] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 6

[40] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 5

[41] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *CVPR*, pages 13294–13304, 2025. 6

[42] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 5

[43] Tang Xu, Wenbin Wang, and Alin Zhong. Hoiedit: Human–object interaction editing with text-to-image diffusion model. *The Visual Computer*, pages 1–13, 2025. 1, 2, 6

[44] Frederic Z. Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring predicate visual context in detecting human–object interactions. In *ICCV*, pages 10411–10421, 2023. 5, 2

[45] Hong Zhang, Zhongjie Duan, Xingjun Wang, Yingda Chen, and Yu Zhang. Eligen: Entity-level controlled image generation with regional attention. *arXiv preprint arXiv:2501.01097*, 2025. 2, 3, 6, 8, 1

[46] Dewei Zhou, You Li, Fan Ma, Zongxin Yang, and Yi Yang. Migc++: Advanced multi-instance generation controller for image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 6

[47] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6818–6828, 2024. 2, 6