OneHOI: Unifying Human-Object Interaction Generation and Editing

Supplementary Material

A. Implementation Details

We build our model by adapting Flux.1 Kontext [21], Eligen [45] and Flux.1 Dev [20] backbone. The text encoder weights are kept frozen during training, and we applied LoRA [15] fine-tuning on the linear layers of each block in the DiT, with a rank of 64, resulting in 0.3 billion trainable parameters (2.5% of the frozen 12B base model). The HOI Encoder (17M) is trained from scratch, while the backbone is adapted via 344M trainable LoRA parameters. We train our model on two NVIDIA RTX 6000 ADA GPUs, with constant learning rate of 1×10^{-4} and bf16 precision. We train on resolution buckets, randomly sampling from the following resolutions (height, width) at each step: (1024, 1024), (768, 1360), (1360, 768), (880, 1168), (1168, 880), (1248, 832), and (832, 1248). For the editing task, we follow Flux.1 Kontext [21] to separate the source image from the noisy latent. The VAE-encoded source image latent patches are assigned RoPE indexes of $(1, x, y)$ while the noise latents are assigned $(0, x, y)$, respectively. For the arbitrary shape, the Fourier embedding $e_{\text{box}}(b_n^r)$ is obtained using the shape's minimum enclosing bounding box. During inference, we use 28 sampling steps and set the classifier-free guidance scale [11] to 3.5.

A.1. Sequence length and budgeting.

Each HOI interaction yields *role sequences* of HOI tokens: subject $\mathcal{S}_n$ object $\mathcal{O}_n$ and action $\mathcal{A}_n$); an object-only case contributes just $(\mathcal{O}_n)$. We cap the total HOI-token budget at K_{HOI} (default to 4608 for 48GB GPU memory) and per-sequence length at $L_{\max}$ (default 512). Let M be the number of active role sequences, we assign the same length L to every active sequence,

$$L = \min\Big(L_{\max}, \ \Big\lfloor \tfrac{K_{\text{HOI}}}{M} \Big\rfloor\Big),$$

so that the total HOI-token count satisfy $M\,L \leq K_{\text{HOI}}$. Practically, each role sequence is padded or truncated to length L for batching.

A.2. Nano Banana.

We compare our method against Nano Banana as a representative closed-source baseline. We access the model via the Gemini API [*] using the `gemini-2.5-flash-image` variant. For fairness, we employ the identical text prompts and source images defined in our editing task (Fig. 6). Since the Gemini API does not currently expose parameters for seed control or stochasticity, we report results from a single inference trial per prompt to evaluate its default zero-shot performance.

[*] https://aistudio.google.com/

A.3. InteractEdit + InteractDiffusion Baseline

To establish a rigorous baseline for layout-guided HOI editing, we integrate the state-of-the-art InteractEdit [14] and InteractDiffusion [13] frameworks. We adapt the original SDXL-based InteractEdit backbone to the InteractDiffusion-XL variant. Our implementation follows a two-stage inversion process for each source image in the IEBench benchmark. In a departure from the standard text-only inversion used in InteractEdit, we leverage InteractDiffusion's native support for structural guidance by incorporating HOI triplets and bounding boxes throughout the inversion stages. Specifically, we execute the inversion for 1000 steps in Stage 1 and 200 steps in Stage 2, adhering to the default configurations of InteractEdit. During the editing phase, we synthesize the final image by conditioned generation using the inverted weights and a structured prompt: "a photo of ⟨subject⟩ ⟨target action⟩ ⟨object⟩ at ⟨background⟩". This process is further guided by the target HOI triplet and the specified HOI layout, ensuring the baseline is evaluated under identical conditioning to our proposed method. Finally, we apply the standard IEBench evaluation strategy to ensure a fair and consistent comparison across all reported metrics.

B. Dataset Details

B.1. Synthesis of Target Layouts for IEBench

The IEBench benchmark [14] is designed for layout-free editing and thus does not provide target bounding boxes for edits. First, we built a statistical geometry bank from the HICO-DET training set. For each HOI class ⟨action,object⟩, we computed a 5-dimensional **multivariate Gaussian distribution**. This distribution models the object's geometry relative to the subject, using a 5D vector that captures the relative centre displacement (dx, dy), relative object size (rw, rh), all scaled by the subject's height, and the Intersection-over-Union (IoU).

To generate a target layout for a specific edit in IEBench, we used this statistical model along with a manually specified heuristic. We categorised objects as *"large/stable"* (*e.g.*, bed, bus) or *"small/movable"* (*e.g.*, skateboard, cell phone). For edits involving **large objects**, we fixed the object's bounding box (b_o) from the source image and sampled a new subject box (b_s) from the learned relative distribution. Conversely, for **small objects**, we fixed the subject's box (b_s) and sampled a new object box (b_o). In some ambiguous cases, both boxes were sampled.

We generated proposals for all 100 edits in IEBench.

These proposals were then manually inspected to filter out any implausible layouts, such as those with unreasonable aspect ratios, sizes, or positions.

B.2. MultiHOIEdit

To evaluate the novel task of multi-HOI editing, for which no existing benchmark exists to our knowledge, we introduce **MultiHOIEdit**. The process began by creating a set of high-quality source images. We used the Flux.1 model to synthesise images containing two or three distinct HOIs, focusing on scenes with different objects to ensure complexity. The generation of plausible multi-HOI images proved to be exceptionally challenging; to ensure correctness, we verified each synthesised image using the PViC HOI detector [44] and retained only those where all target interactions were successfully detected. This rigorous filtering process had a very low yield, with only **200 valid source images** being selected from an initial pool of 8,942 generations (a 2.2% success rate), underscoring the difficulty of the task.

From this curated set of source images, we then defined the target edits. For each source image, we created one to three distinct editing tasks, where each task involved modifying two or more of the existing HOIs simultaneously. The target layouts for these new interactions were proposed by extending the statistical geometry bank method described in Appendix B.1 and were then manually filtered for quality and plausibility. The final MultiHOIEdit benchmark comprises **103 unique source images** and a total of **200 distinct multi-interaction editing tasks**. Qualitative examples of these complex edits are provided in Fig. 24.

The benchmark is diverse, covering **54 object categories** (Fig. 18) and a total of **40 source actions** (Fig. 19) and **74 target actions** (Fig. 20). Overall, the tasks involve transitions between **112 source HOI-object pairs** and **252 target HOI-object pairs**, with the full range of edits detailed in Fig. 21. We will release **MultiHOIEdit** publicly.

B.3. HOI-Edit-44K

The HOI-Edit-44K dataset addresses the critical scarcity of large-scale, paired data for the task of human-object interaction editing. The final dataset consists of 44,117 high-quality, paired HOI editing examples. Each sample in the dataset includes (1) the source image, (2) the target interaction triplet (subject, object, action), (3) the edited image and (4) the corresponding HOI layout for the edited image.

The dataset is diverse, containing 79 unique object categories (Fig. 16) and 92 unique target actions (Fig. 17), which combine to form 372 unique HOI triplets. See Fig. 15 for qualitative examples. This resource was critical for jointly training our unified model, providing the necessary supervision for robust, identity-preserving HOI editing .

Generalization and reliability. Identity-preserving HOI edit pairs are scarce, necessitating our strictly curated HOI-Edit-44K. Our sources images are not purely synthetic as they come from both Flux.1 generations and real HICO-DET photos. We retain pairs only if they satisfy two rigorous criteria: HOI correctness via PViC and identity consistency via DINOv2 ($\geq$ 0.75). This strict quality control yields a $\sim$90% rejection rate, which ensures the high reliability and physical plausibility of the final 44K curated pairs. Crucially, we also jointly train on HOI generation using real HICO-DET images. This exposes the model to real-scene statistics and interaction distributions beyond synthetic edits, anchoring the learned representation in real-image distributions and effectively mitigating potential teacher-model bias.

C. Evaluation

C.1. Interaction Editing

For Interaction Editing task, we follow the evaluation protocol of InteractEdit [14] in their proposed IEBench. These evaluation metrics specifically designed for HOI editing task and 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 [44], 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) Editability-Identity Score, EI quantifies the trade-off between HE score and Identity Consistency via the harmonic mean, analogous to the F_1 score [32]. This formulation ensures a balanced evaluation by penalizing low performance in either dimension:

$$\text{EI} = \frac{2 \times \text{HOI Editability} \times \text{Identity Consistency}}{\text{HOI Editability} + \text{Identity Consistency}}. \quad (5)$$

Here, Identity Consistency assesses how well the subject and object identities are preserved after editing. To compute this, GroundingDINO [25] and SAM [18] is used to detect and segment the subject and object in both the source and edited images. Then, DINOv2 [29] is used for extracting feature embeddings, and the cosine similarity between embeddings of source-subject and edited-subject (and similarly for the object) is calculated and aggregated over images and seeds.

C.2. Human Evaluation Study

To complement the quantitative results, we conducted a rigorous human preference study evaluating **HOI Correctness**, **Identity Preservation**, and **Overall Quality**. The study utilized a blind, randomized side-by-side comparison format where 26 unique respondents evaluated a total

of **450 trials** (N=450). As illustrated in the provided survey interface in Fig. 11, participants were presented with a source image and a specific edit instruction, such as "Make the person ride the skateboard". For each trial, respondents rated two anonymized outputs: our model versus a baseline, using a 5-point Likert scale ranging from "A much better" to "B much better," with an "Equal" option for ties.

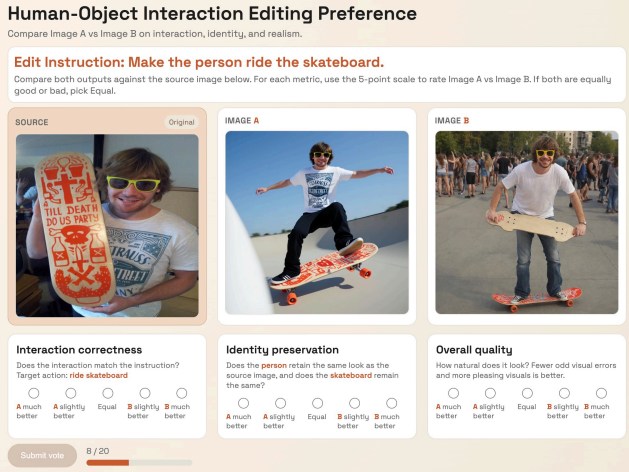

Figure 11. Evaluation Interface. Web-based survey used for data collection. Participants performed side-by-side comparisons of two models based on a source image and target edit instruction.

The results, summarized in Fig. 12, demonstrate that our method significantly outperforms leading baselines in physical plausibility and structural coherence. When compared against QwenImageEdit, our model was preferred in 58.2% of cases for HOI Physics Plausibility, while the baseline was favoured in only 8.2% of trials. Furthermore, our approach achieved a commanding 72.0% win/tie rate in Overall Quality, consisting of a 50.4% outright win rate and a 21.6% tie rate. In comparisons with InteractEdit, our model maintained a superior win rate for Identity Preservation (74.8%) and Overall Quality (66.1%). These findings suggest that our unified representation effectively resolves the trade-off between executing complex interaction edits and maintaining the structural identity of the original scene.

D. Additional Qualitative Results

We provide additional qualitative results for the layout-free HOI editing task in Figure 23. Likewise, Figure 22 presents additional qualitative results for HOI generation. Furthermore, Figure 14 serves as a visual representation to the paper's core question, demonstrating that **HOI generation and editing are successfully unified within a single framework.** The step-by-step workflow showcases the seamless integration of initial HOI generation, multi-HOI editing, single-HOI editing, and attribute editing, thereby demonstrate the comprehensive and versatile control enabled by our method.

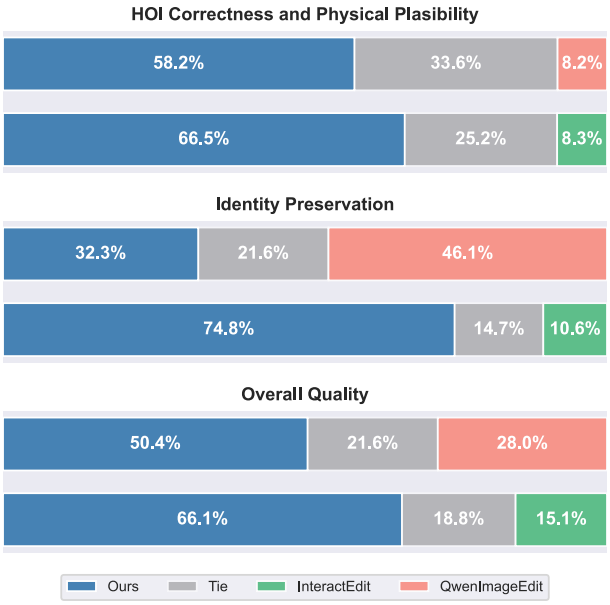

Figure 12. Results of the Human Preference Study. Aggregated preference percentages for HOI Correctness and Physical Plausibility, Identity Preservation, and Overall Quality. The top bar in each category compares OneHOI (Ours) against QwenImageEdit, while the bottom bar compares it against InteractEdit.

Spatial action region for remote action. We use subject ∪ object as an attention-aligned action grounding prior. Fig. 4 (main paper) shows that for disjoint interactions, action-token attention concentrates on entities, and *union* matches this footprint better than the "*Between*" band. We further validate this on a trajectory verb ("throwing frisbee" in Fig. 13). The action-token attention focuses on the thrower and frisbee, and the *union* region matches this footprint, while the "*Between*" band is often narrow/misplaced.

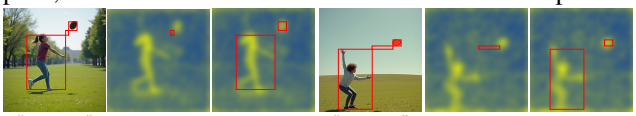

Figure 13. Attention footprint of Flux.1. "Union" better matches the attention footprint compared to "Between".

E. Ablation on Unification vs. Task-specific

We unify HOI generation and editing by supporting mixed conditioning for real-world use cases (text-only, partial layouts, or multi-HOI). Separate training yields brittle, task-specific priors. Notably, the generation becomes strictly layout-dependent, while editing fails to scale to multi-HOI. As shown in Tab. 5, **Unified** model consistently outperforms **task-specific** models trained under matched computation (1k steps), improving HOI Accuracy by 26.4% in generation and HOI Editability by 21.1% in layout-free editing. This confirms that joint training enables a *"synergy effect"*, where generative priors enhance editing robustness and vice versa. (*Note*: Task-specific = single-task models)

Table 5. Ablation on Unification.

Task Scenario	Metric	Task-specific	**Unified (Ours)**
Generation	Spatial ↑ / HOI Acc. ↑	0.422 / 0.177	**0.443 / 0.224**
Layout-Free Edit	Editability-Identity ↑ / HOI Editability ↑	0.574 / 0.464	**0.611 / 0.562**
Multi-HOI Edit	Editability-Identity ↑ / HOI Editability ↑	0.391 / 0.287	**0.435 / 0.329**

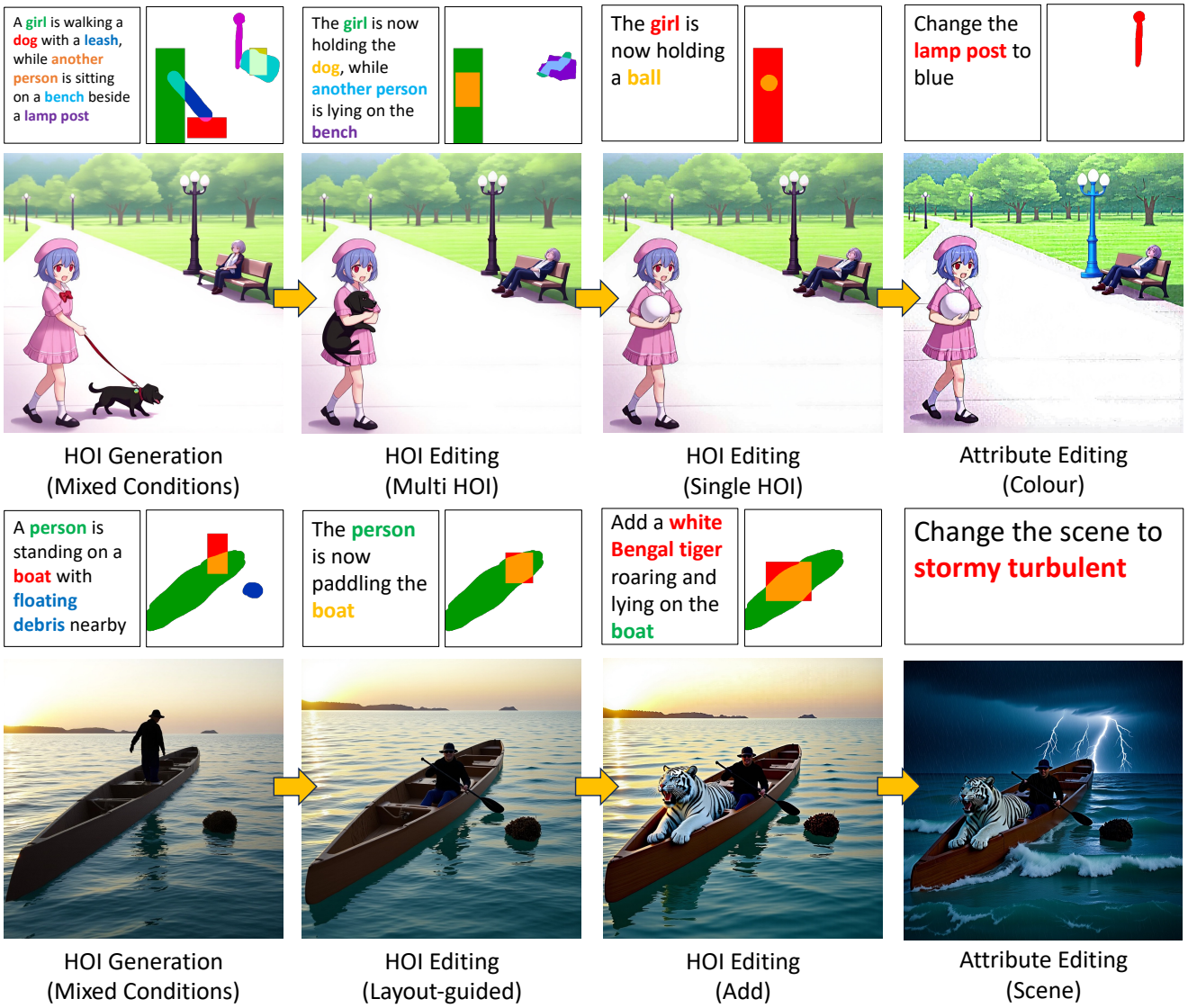

Figure 14. **Versatile workflow for unified HOI generation and editing using OneHOI. OneHOI** enables a seamless, multi-step workflow within a single model, showcasing diverse conditional control. Starting with:
Top Row: Urban Park Scene. (1) **Mixed-Condition Generation** synthesises a complex scene from layout-guided HOIs (*i.e.*, *walking dog*) and arbitrary shape-guided independent objects (*i.e.*, *lamp post*, *leash*), alongside another HOI (*i.e.*, *person sitting on bench*). (2) **Multi-HOI Editing** simultaneously updates two distinct interactions (*i.e.*, *holding dog*, *person lying on bench*). (3) **Single-HOI Editing** modifies one interaction (*i.e.*, *holding ball*). (4) **Attribute Editing** changes an object's colour (*i.e.*, *black* → *blue*).
Bottom Row: Ocean Survival Scene. (1) **Mixed-Condition Generation** creates a challenging open-water scenario from a person standing on a boat and arbitrary shape-guided floating debris. (2) **Layout-guided HOI Editing** precisely changes the person's action (*i.e.*, *paddling the boat*). (3) **HOI Editing (Add)** introduces a new interaction (*i.e.*, *white Bengal tiger roaring and lying on the boat*). (4) **Attribute Editing (Scene)** transforms the entire environment (*i.e.*, *day* → *stormy*, *calm* → *turbulent* ocean).

Figure 15. Examples from the HOI-Edit-44K dataset.

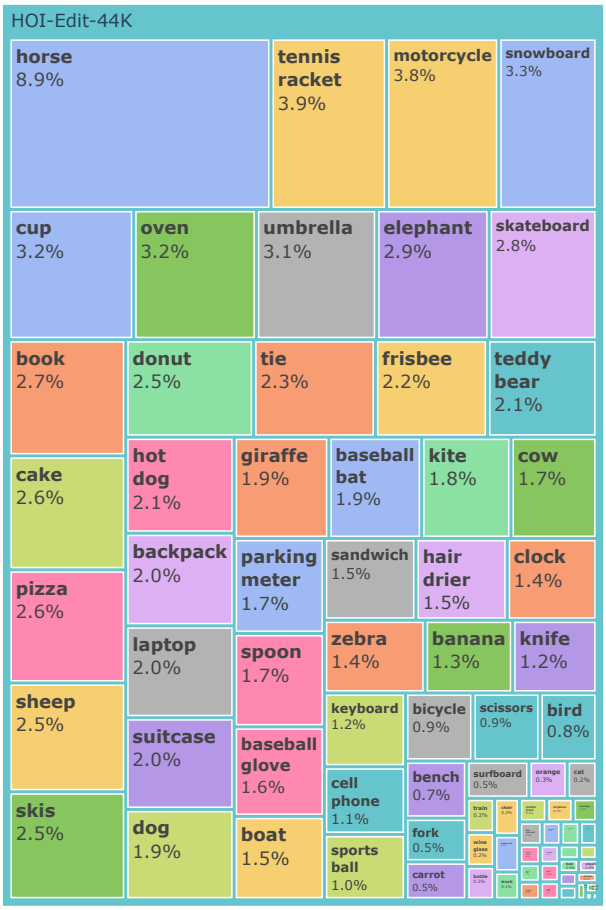

Figure 16. Treemap visualising the distribution of the interacting object categories in the HOI-Edit-44K dataset. The size of each block corresponds to the category's frequency.

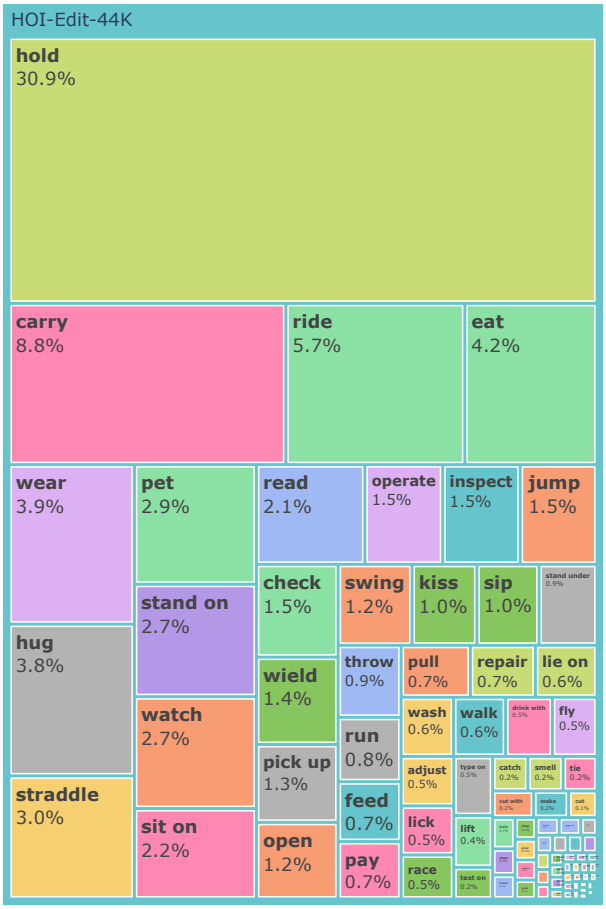

Figure 17. Treemap visualising the distribution of target action categories in the HOI-Edit-44K dataset.

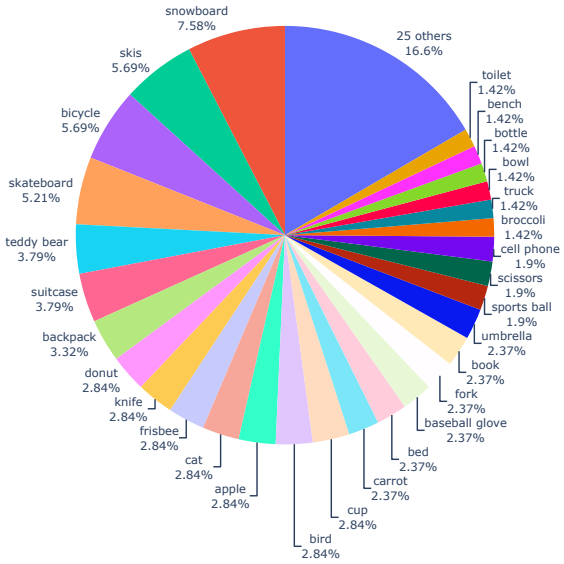

Figure 18. Distribution of the 54 object categories within the MultiHOIEdit. The "25 others" aggregates the least frequent categories with 2 or fewer appearance.

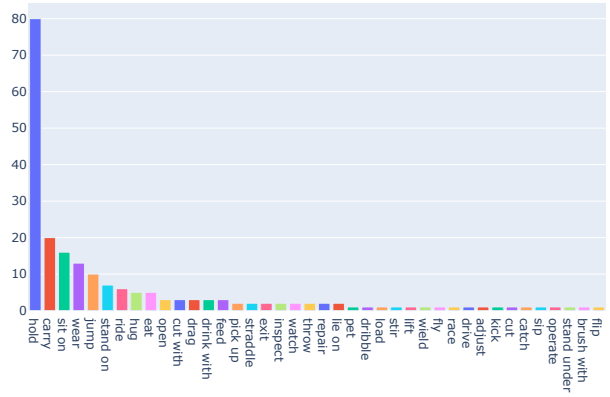

Figure 19. Distribution of source (pre-edit) actions in MultiHOIEdit.

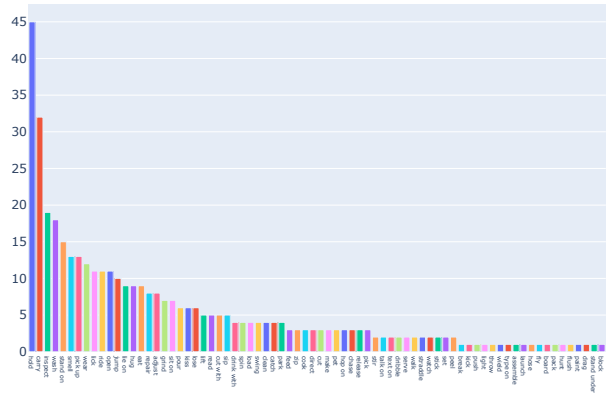

Figure 20. Distribution of 74 target (post-edit) actions in MultiHOIEdit.

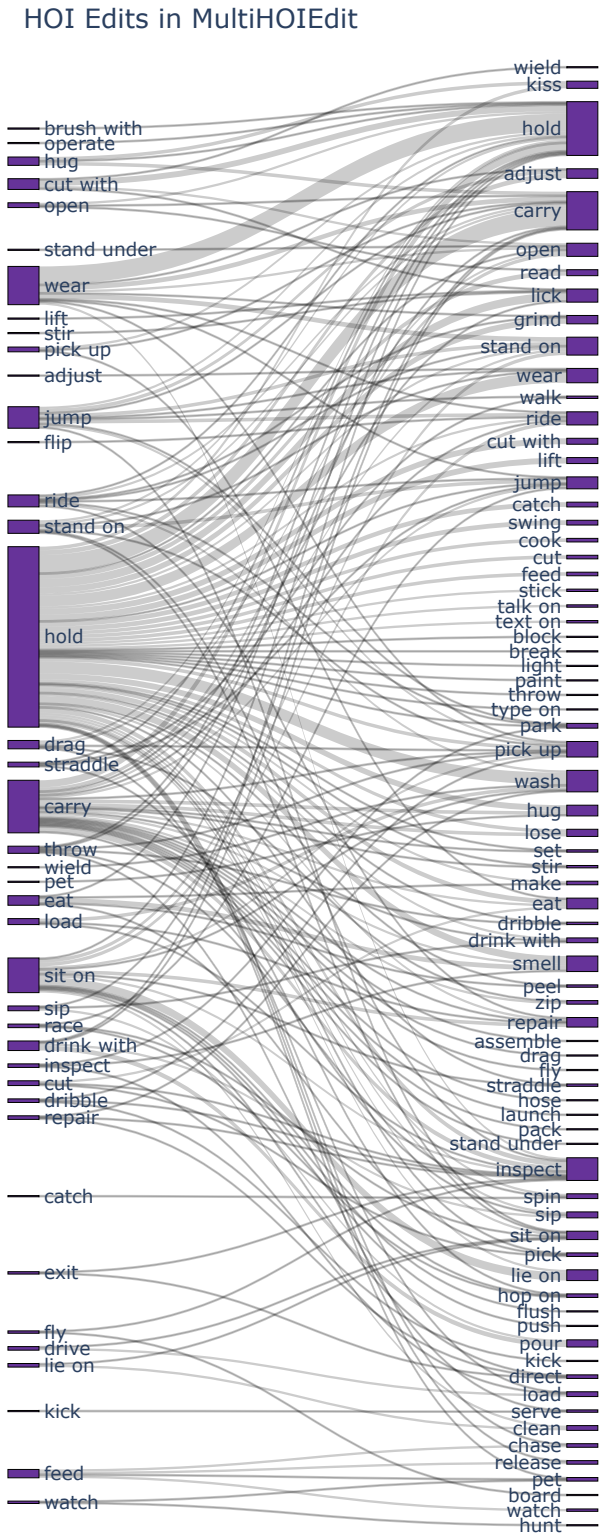

Figure 21. Sankey diagram visualising the action transitions in the MultiHOIEdit benchmark. The flows illustrate the mapping from source actions (left) to target actions (right), detailing the full range of edits.

Figure 22. Additional qualitative results for HOI generation. These examples further highlight the limitations of baselines, which often fail to render the specific action even when the objects are placed correctly. In the **first row (standing on a chair)**, all baseline methods incorrectly generate a child sitting on a chair, while our model is the only one that correctly synthesises the 'standing on' pose. Similarly, for **holding a spoon (row 2)**, baselines produce general eating scenes, with Eligen and InteractDiff showing a fork instead. Our model, in contrast, correctly renders the person holding a spoon. This challenge is more pronounced in complex multi-HOI prompts. For **row 3 (flipping, jumping, and riding a skateboard)**, baselines fail to capture the 'flipping' or 'jumping' motions, rendering a simple 'riding' pose at best. In **row 4 (drinking with bottle while holding it)**, most methods fail to combine both 'holding' or 'drinking'. In contrast, our model generates coherent images that plausibly reflects all specified interactions, demonstrating superior compositional understanding.

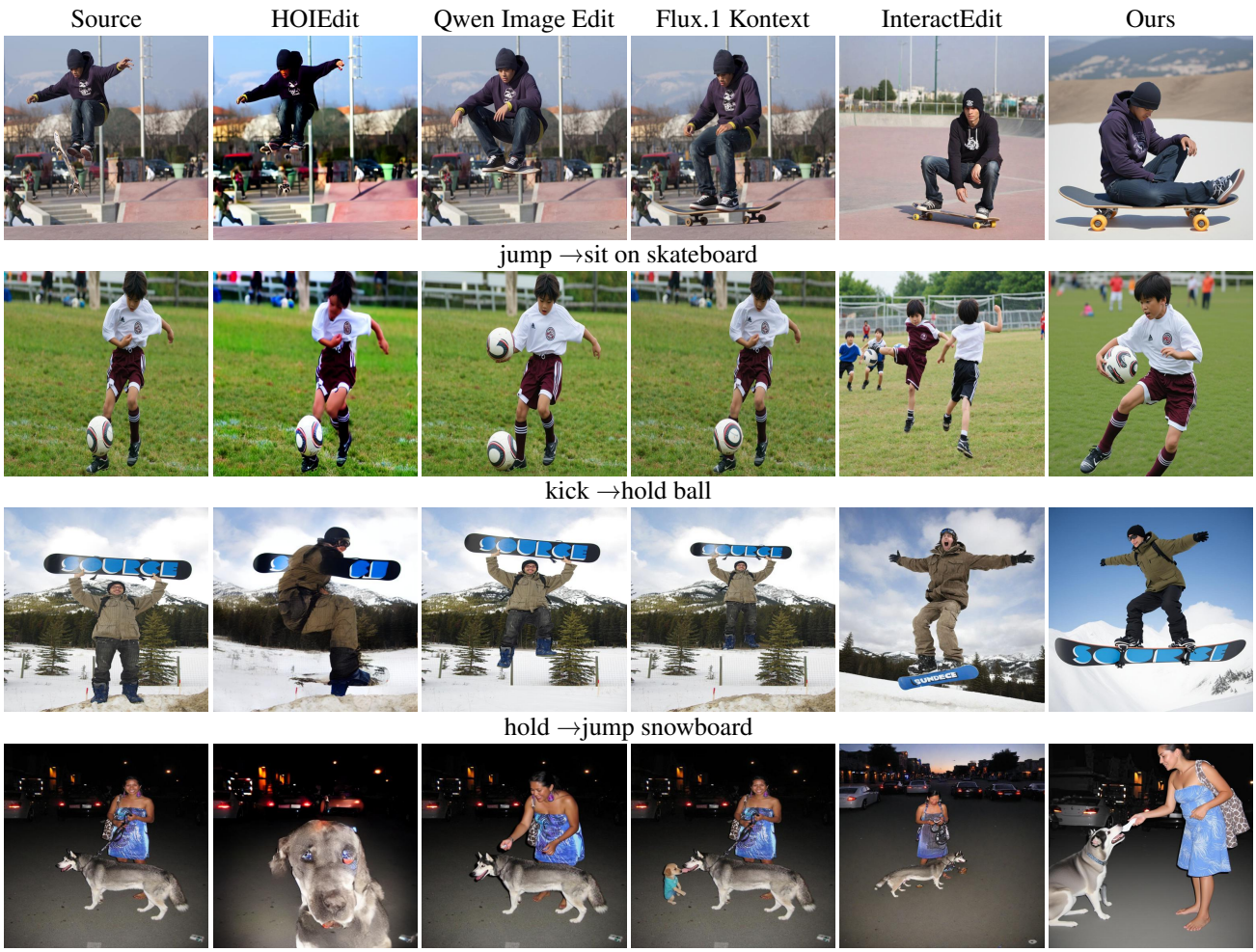

Figure 23. Additional qualitative comparisons for layout-free HOI edits. **Row 1 (jump → sit on skateboard)**: Baselines show incorrect poses (Qwen, Flux.1), unnatural actions (InteractEdit), or severe artifacts (HOIEdit), while ours renders "sit on" interaction. **Row 2 (kick → hold ball)**: Most baselines fail to alter the pose, while ours renders the "hold" action. **Row 3 (hold → jump snowboard)**: Most methods failed to render "jump". Although InteractEdit renders jump, it fails to preserve the snowboard's identity. Ours renders the jump while maintaining the identity of both the person and the snowboard. **Row 4 (walk → feed dog)**: Only ours renders a coherent "feeding" interaction while preserving the identities of both subjects, demonstrating its superior capability in handling complex relational changes.

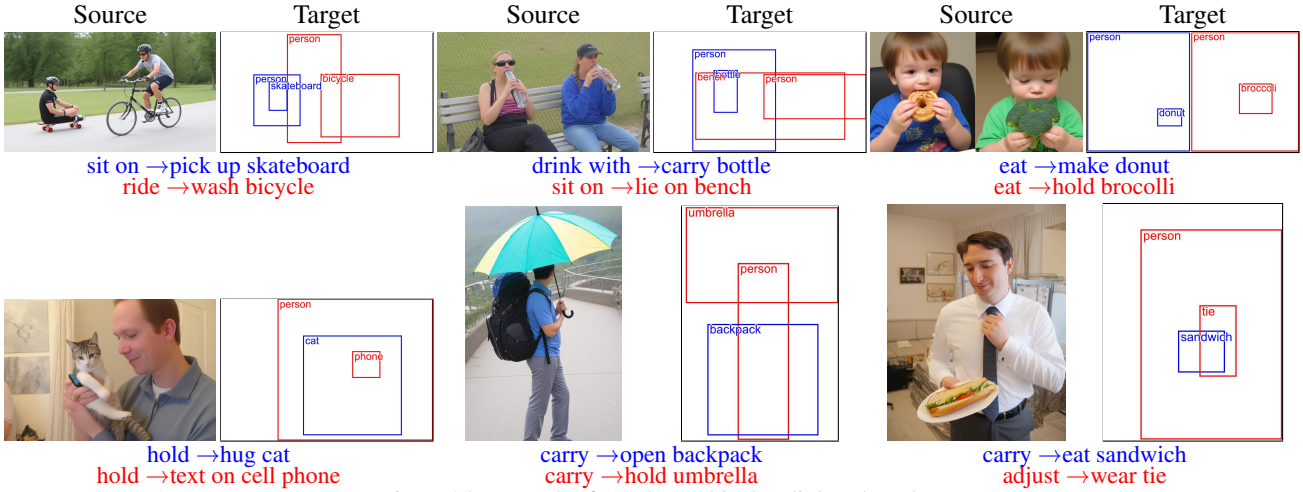

Figure 24. Examples from the MultiHOIEdit benchmark.