OneHOI: Unifying Human-Object Interaction Generation and Editing

Jiun Tian Hoe · Weipeng Hu · Xudong Jiang · Yap-Peng Tan · Chee Seng Chan

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE · SUN YAT-SEN UNIVERSITY · UNIVERSITI MALAYA · VINUNIVERSITY

CVPR JUNE 3-7, 2026 DENVER COLORADO

TL;DR

Can HOI generation and editing be unified?
YES **-** and they actually make each other better.

Motivation

Human-Object Interaction (HOI)
= <person, action, object>

Current Limitations in HOI content creation

HOI Editing

- Only text-driven edits
- Fails to decouple pose from physical contact
- Limited to single HOI
- Lack spatial control
- Limited HOI semantic understanding

HOI Generation

- Hard to accept arbitrary shape
- Cannot mix HOI with object-only entities
- Input conditions not flexible

Key Insight

Generation and editing are the same conditional denoising process

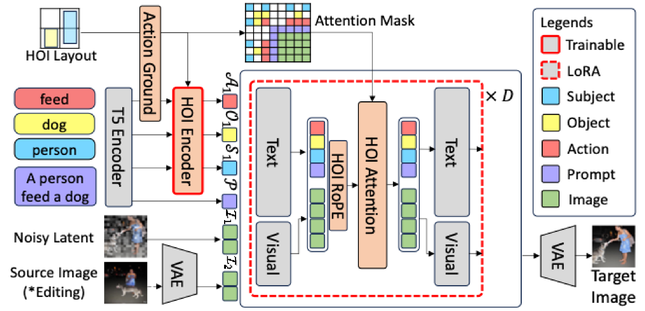
Joint learning creates strong **synergy**; broad semantics from generation enable more plausible and physically-aware edits.

Our Contributions

- First unified framework for HOI generation & editing
- Relational Diffusion Transformer (R-DiT) with explicit interaction modeling
- Supports layout-guided, layout-free, arbitrary-mask, and mixed condition control
- Enable multi-HOI editing
- HOI-Edit-44K dataset, first for HOI editing pairs
- State-of-the-art result across all HOI synthesis task

OneHOI

Overall Framework



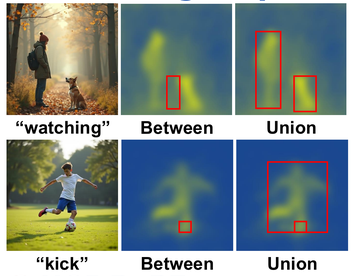
Train Strategy

Joint Training
Generative + Editing

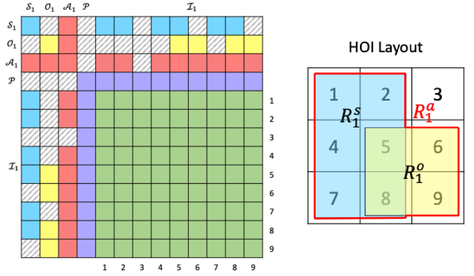
Modality Dropout
Allow flexible condition combinations

Multi-source Data
HOI-Edit-44K and HOI generation data

1 Action Region (Union)

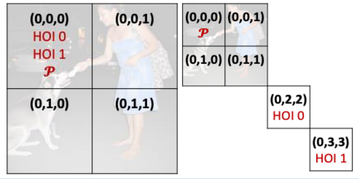


2 Structured HOI Attention

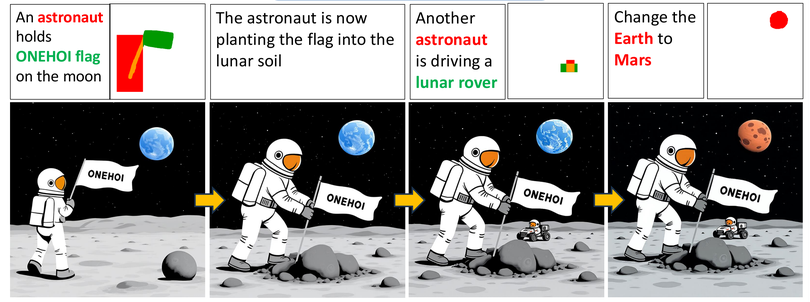


1. Our "union" better aligns with action attention maps in DiT
2. Verb-mediated attention allows explicit interaction modelling
3. HOI RoPE assigns distinct positional slots to distinguish multiple HOIs

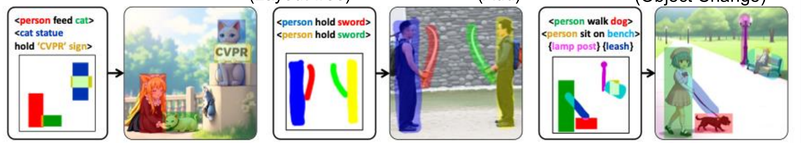
3 HOI RoPE



Results



HOI Generation · HOI Editing (Layout-free) · HOI Editing (Add) · Attribute Editing (Object Change)



HOI Generation · Arbitrary Shape · Mixed Conditions

HOI Editing · Layout-guided HOI Editing · Multi HOI Editing

Layout-free (Text-only) Editing

Method	HOI Editing		Image Quality		
	Editability-Identity	HOI Editability	PickScore	HPS	ImageReward
Null-Text Inversion	0.443	0.390	20.81	0.2483	-0.3329
InstructPix2Pix	0.380	0.269	20.28	0.2178	-0.7717
Flux.1 Kontext	0.471	0.328	20.45	0.2427	-0.5137
Qwen Image Edit	0.580	0.460	20.81	0.2585	0.0748
InteractEdit	0.573	0.514	21.08	0.2640	0.1630
Nano Banana (close)	0.623	0.530	20.97	0.2544	0.1810
Ours	**0.638**	**0.596**	**21.26**	**0.2805**	**0.4713**

HOI Generation

Method	Controllability		Image Quality		
	Spatial	HOI	PickScore	HPS	ImageReward
GLIGEN	0.5150	0.3344	20.46	0.2322	−0.4103
InstanceDiffusion	0.5228	0.3476	20.06	0.2312	−0.2532
MIGC++	0.5331	0.3616	20.16	0.2208	−0.6492
Eligen	0.4371	0.3061	21.28	0.2496	0.3921
InteractDiffusion	0.5768	0.4505	20.37	0.2283	−0.3194
Ours	**0.6104**	**0.4528**	**21.41**	**0.2617**	**0.5224**

Layout-guided HOI Editing

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