Taming I2V models for Image HOI Editing: A Cognitive Benchmark and Agentic Self-Correcting Framework

Jiayi Gao[1], Qingchao Chen[2], Yuxin Peng[1], Yang Liu[1]

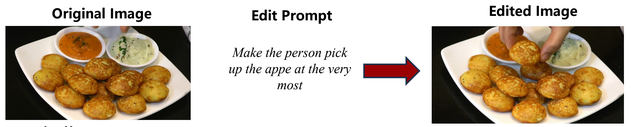
1 Wangxuan Institute of Computer Technology, Peking University 2 National Institute of Health Data Science, Peking University

Code Paper

Introduction

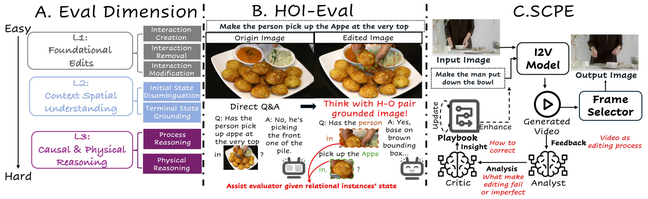
Task: Human-Object Interaction (HOI) Image Editing

- **Inputs:** Original Image + Edit Prompt
- **Outputs:** Image preserving the original content and matching human-object interactions in the edit prompt.



Core Challenge

- **No Benchmark:** Need basic interaction & high-level reasoning in HOI editing.
- **Weak Metrics:** Need region-sensitive pair assessment.
- **Dynamic Remodeling:** Need interaction remodeling, not static visual appearance.



HOIEdit Dataset

HOIEdit: A **benchmark** assessing across **3 progressive cognitive tiers.**

- **L1 Foundational Editing:** Edit basic human-object interactions.
- **L2 Spatial Understanding :** Ground interactions in right objects and places.
- **L3 Causal & Physical Reasoning:** Ensure logical steps and physical effects.

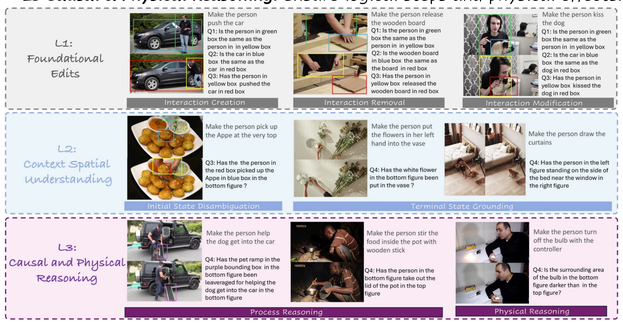


Figure 2. **Illustrative examples for hierarchical cognitive level in *HOIEdit*.**

Data Curation: Fine-grained HOI evaluation data.

- Scene-aware editing instructions
- Human / object / context region annotations
- Questions for identity, interaction, space, and reasoning

HOI-Eval: Region-grounded HOI assessment.

- **Localize:** track target human-object regions
- **Verify:** compare human/object identity
- **Assess:** VLM judges interaction, spatial grounding, and reasoning ability

Figure 4. HOI-Edit data distribution

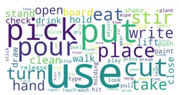


Figure 5. Interaction verbs

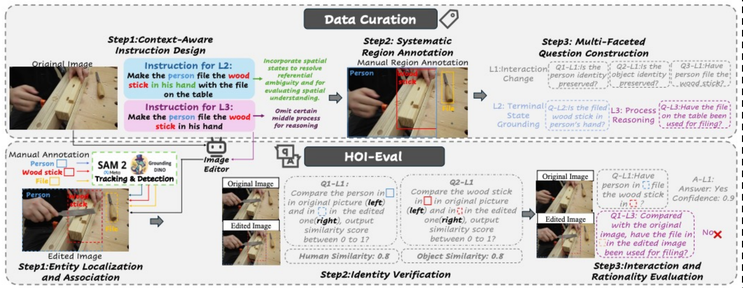
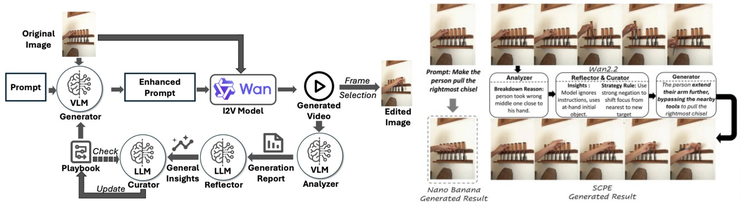


Figure 3. **Overview of benchmark construction and evaluation pipeline.**

Self-Correcting Process Editing

- **SCPE:** Self-correcting HOI editing with I2V traces.
  - **Generator:** refine prompt with Playbook
  - **Analyzer:** detect frame-level failures
  - **Reflector:** turn failures into insights
  - **Curator:** update reusable strategies
  - **Output:** select the best edited frame
- **Dynamic Playbook:** Reusable memory for prompt correction.
  - Stores strategies, pitfalls, and templates
  - Learns from failure cases
  - Guides precise HOI prompt generation



Experiments

Quantitative Results

Table 1. **Quantitative Comparison.** We report Interaction (I), Human (H), and Object (O) consistency across three cognitive levels. Note that **I+Q&A** sets the interaction score to 0 if the answer for context-aware question is wrong. SCPE only update for 1 iteration.

Method	Source	L1: Foundational			L2: Understanding				L3: Reasoning			
		I↑	H↑	O↑	I↑	I+Q&A↑	H↑	O↑	I↑	I+Q&A↑	H↑	O↑
Flux.1 Kontext (Esser et al., 2024)	Open	0.4961	0.8991	0.5347	0.5192	0.2780	0.8998	0.4631	0.2052	0.0423	0.9653	0.8777
ByteMorph (Chang et al., 2025)	Open	0.4469	0.2000	0.2220	0.4279	0.2240	0.2227	0.1757	0.4156	0.1571	0.7480	0.6584
Step1X-Edit(Liu et al., 2025b)	Open	0.5396	0.7985	0.6533	0.5159	0.4058	0.7997	0.6472	0.5874	0.4194	0.8446	0.7640
ChronoEdit(Wu et al., 2025b)	Open	0.5823	0.7418	0.6023	0.5574	0.4608	0.7515	0.6160	0.5620	0.4345	0.8205	0.7359
Bagel(Deng et al., 2025)	Open	0.6326	0.7940	0.4790	0.6065	0.4781	0.7804	0.5030	0.6013	0.4061	0.8516	0.5701
Qwen-Image-Edit PLUS(Wu et al., 2025a)	Closed	0.6128	0.9343	**0.8775**	0.5984	0.4928	0.9395	0.7924	0.5878	0.3870	0.9593	0.8602
Nano Banana(Google, 2025)	Closed	0.7271	**0.9537**	0.8609	0.7040	0.5960	**0.9590**	0.7706	0.7399	0.5782	**0.9743**	**0.9185**
Wan 2.2 I2V (Wan et al., 2025)	Open	0.6908	0.9166	0.7823	0.6608	0.5526	0.9113	0.7272	0.6822	0.5343	0.9306	0.8511
Wan 2.2 I2V + SCPE	Open	**0.8423**	0.9260	0.8640	**0.7909**	**0.6952**	0.9269	**0.8260**	**0.8053**	**0.6528**	0.9518	0.9073

Table 4. **Quantitative Comparison on SCPE on different variants with inference time.** SCPE only updates for one iteration.

Method	Time (min)	Overall		
		I↑	H↑	O↑
ChronoEdit(Wu et al., 2025b)	20.0	0.5709	0.7607	0.6337
Nano Banana(Google, 2025)	-	0.7231	0.9595	0.8469
+ SCPE (Ours)	-	0.7310	**0.9602**	0.8351
Wan 2.2 I2V(Wan et al., 2025)	30.0	0.6804	0.9189	0.7807
+ SCPE (Ours)	30.0	**0.8199**	0.9267	**0.8565**
TurboDiffusion(Zhang et al., 2025a)	**1.4**	0.6290	0.8890	0.6967
+ SCPE (Ours)	**1.4**	0.7536	0.9118	0.7735

Qualitative Results

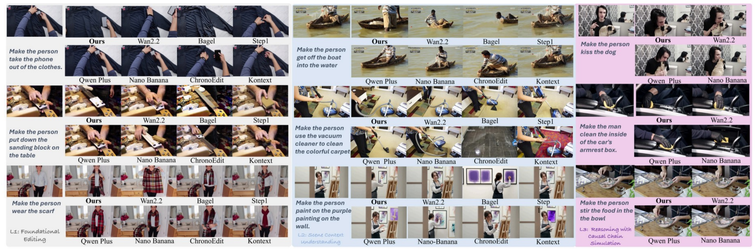


Figure 6. **Qualitative comparison on HOI-Edit.** where 'Qwen Plus' stands for Qwen-Image-Edit PLus

Metrics Reliability

Table 2. **Correlation with Human Judgment.** We report the Pearson correlation (*Pr*) and statistical significance (*Pp*).

Method	Subject (H)		Object (O)		Interaction	
	Pr	Pp	Pr	Pp	Pr	Pp
DINOv2	0.035	0.835	0.172	0.301	-	-
CLIP	0.253	0.125	-0.027	0.872	-	-
HOIEval (Ours)	**0.43**	**.007**	**0.47**	**.003**	**0.60**	**.002**

Figure 9. Subject-object similarity comparison: HOI-Eval vs global metrics (instruction: place jacket on car hood).

Robustness Analysis

Table 5. Quantitative comparison on HOI-Edit benchmark. Models and results are processed under independent evaluator GPT-5.4.

Method	Source	L1 (Foundational)			L2 (Understanding)			L3 (Reasoning)		
		I↑	H↑	O↑	I+Q&A↑	H↑	O↑	I+Q&A↑	H↑	O↑
Nano Banana	Closed	0.7346	**0.9437**	**0.9195**	0.6134	**0.9391**	0.881	0.5587	**0.9437**	**0.9319**
Wan 2.2	Open	0.7325	0.8865	0.8804	0.6104	0.8978	0.8033	0.5012	0.904	0.8832
Wan 2.2 + SCPE	Open	**0.7870**	0.8999	0.8967	**0.6595**	0.913	**0.8928**	**0.5878**	0.9218	0.9066

Table 6. **Evaluation of VLM-agnostic flexibility.** This comparison isolates the Playbook's logic from the specific reasoning engine, comparing Gemini and Qwen 3.5-VL. The near-consistent gains indicate that the accumulated physical priors, rather than the inherent strength of a specific VLM, are the primary drivers of dynamic editing performance.

Method	Source	L1 (Foundational)			L2 (Understanding)				L3 (Reasoning)			
		I↑	H↑	O↑	I↑	I+Q&A↑	H↑	O↑	I↑	I+Q&A↑	H↑	O↑
Wan 2.2 I2V (Wan et al., 2025)	Open	0.6908	0.9166	0.7823	0.6608	0.5526	0.9113	0.7272	0.6822	0.5343	0.9306	0.8511
Wan 2.2 I2V + SCPE	Open	**0.8423**	**0.9260**	**0.8640**	0.7909	**0.6952**	**0.9269**	**0.8260**	**0.8053**	**0.6528**	0.9518	**0.9073**
Wan 2.2 I2V + Qwen Playbook	Open	0.8368	0.9115	0.8427	**0.7922**	0.6857	0.9048	0.7724	0.7928	0.6454	**0.9551**	0.8835
Wan 2.2 I2V + Gemini OPE	Open	0.7396	0.8287	0.6658	0.7194	0.5910	0.8115	0.6429	0.6988	0.5724	0.8326	0.7013

- Consistent gains across evaluators and VLM agents.
- Proving improvements come from video-feedback correctionand Playbook memory, not evaluator bias or a specific VLM.

Ablation study

Table 3. **Ablation study. I** is interaction score and IDS is identity score (avergae of **H** & **O**).

Method Variant	I↑	IDS↑
Wan 2.2 I2V 14B	0.6804	0.8494
+ OPE	0.7028	0.7385
wo Playbook	0.7625	0.8786
+ SCPE	**0.8199**	**0.8954**

- SCPE performs best on both I and IDS.
- OPE gives small interaction gains but damages identity.
- Playbook memory is key to stable correction.

Playbook Visualization