

UniHOI: Unified Human-Object Interaction Understanding via Unified Token Space

Panqi Yang^{1*}, Haodong Jing^{1*}, Nanning Zheng¹, Yongqiang Ma^{1†}

¹State Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center of Visual Information and Applications, and Institute of Artificial Intelligence and Robotics, Xi’an Jiao Tong University
{yangpq,jinghd}@stu.xjtu.edu.cn, {nanzheng,musayq}@xjtu.edu.cn

Abstract

In the field of human-object interaction (HOI), detection and generation are two dual tasks that have traditionally been addressed separately, hindering the development of comprehensive interaction understanding. To address this, we propose UniHOI, which jointly models HOI detection and generation via a unified token space, thereby effectively promoting knowledge sharing and enhancing generalization. Specifically, we introduce a symmetric interaction-aware attention module and a unified semi-supervised learning paradigm, enabling effective bidirectional mapping between images and interaction semantics even under limited annotations. Extensive experiments demonstrate that UniHOI achieves state-of-the-art performance in both HOI detection and generation. Specifically, UniHOI improves accuracy by 4.9% on long-tailed HOI detection and boosts interaction metrics by 42.0% on open-vocabulary generation tasks.

Introduction

The Human-Object Interaction (HOI) understanding encompasses both HOI detection (identifying ⟨human, action, object⟩ triplets in images) and HOI generation (synthesizing images conditioned on specified interactions). These two tasks are essentially inverse to each other, sharing highly related underlying semantic representations and reasoning processes. However, existing approaches largely treat detection and generation as isolated problems: most HOI generation methods (Li et al. 2023b; Hoe et al. 2024; Gao et al. 2020) rely on explicit spatial constraints (e.g., bounding boxes), limiting their ability to generalize to novel or complex interactions. Meanwhile, state-of-the-art HOI detection models (Wang et al. 2024; Luo et al. 2024; Hui et al. 2025) focus solely on recognition and remain decoupled from generation, which hinders knowledge sharing and demands extensive fine-grained annotations for each task. This significantly limits the scalability of HOI understanding in open-world scenarios.

This motivates a key question: *Can we unify HOI detection and generation within a single framework to fully exploit their shared semantic representations?* We contend

*These authors contributed equally.

†Corresponding author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

that such unification would promote richer interaction understanding, improve data and knowledge efficiency, and enable cross-task and open-world generalization, thus paving the way for more holistic HOI reasoning.

The recent emergence of Multimodal Large Language Models (MLLMs) offers a promising approach for unifying perception and generation within a single, expressive framework. Inspired by works like MMaDA (Yang et al. 2025) and particularly Liquid (Wu et al. 2024b), which demonstrated the potential of unified token spaces for joint vision-language modeling, we argue that a unified representation for interaction semantics holds the key to bridging the HOI detection-generation gap. To this end, our approach is grounded in two key insights:

First, **Unified Token Space**: the semantic essence of a HOI triplet ⟨human, action, object⟩ can be effectively represented within the same discrete vocabulary as visual tokens derived from images (e.g., VQGAN (Esser 2020)). This enables bidirectional mapping within a MLLM, where a textual prompt like “person feeding cat” can evoke spatial relationships (“hand-bowl-cat”) during generation, while visual tokens can be decoded back into semantic triplets during detection.

Second, **Dual Complementarity**: HOI detection and generation are inherently complementary; detection provides explicit interaction priors (e.g., typical spatial distributions and fine-grained annotations) to guide generation, while generation encourages the model to acquire richer, more compositional interaction representations. This reciprocal relationship improves both the expressiveness and generalization of the learned HOI representations, allowing the model to effectively acquire unified HOI semantic representations.

Therefore, we propose **UniHOI**, the first unified semi-supervised MLLM for HOI detection and generation. UniHOI operates on a massively expanded vocabulary merging visual codebook tokens and text tokens, accepting either images (for detection) or text prompts (including structured HOI triplets for generation) to produce corresponding outputs. Specifically, our approach incorporates: (1) an **Interaction-Aware Attention (IAA)** module that injects HOI triplet embeddings into cross-attention layers to focus on interaction-relevant regions; (2) a **semi-supervised learning framework** grounded in cycle consistency within



Figure 1: UniHOI is the first to achieve unified modeling for the two inverse tasks of HOI detection and generation. Through a unified token space, our method enables generalizable interaction semantics understanding and cross-task knowledge sharing. UniHOI achieves state-of-the-art results on most metrics for both HOI detection and generation. Here, HICO-D refers to the *Rare* metric of the *Default* split in the HICO-DET (Chao et al. 2017), other abbreviations follow similarly.

the discrete token space, which enables effective joint training of HOI detection and generation even under limited or heterogeneous supervision. By integrating strong supervision (HICO-DET (Chao et al. 2017)), weak supervision (image-text pairs from LAION-SG (Li et al. 2024)), and unlabeled images, this framework mitigates reliance on exhaustive HOI annotations and facilitates the learning of more generalizable interaction representations for both tasks. And our contributions are as follows:

- We demonstrate that interaction-aware semantic representations can be jointly encoded and reasoned about in a unified discrete token space. Our **modality-aware unified token space** enables bidirectional mapping and compositional reasoning between HOI detection and image generation within a single model, surpassing traditional embedding-level alignment approaches.
- We propose an **Interaction-Aware Attention (IAA)** module with parameter-shared, symmetric cross-attention that explicitly encodes structured HOI semantics as relational priors for both detection and generation. This unified mechanism enables interpretable, context-aware cross-modal reasoning.
- We present a unified **semi-supervised learning strategy** based on cycle consistency in the shared token space, allowing effective training with mixed supervision and reducing annotation cost for open-world HOI recognition and generation.
- Extensive experiments demonstrate that UniHOI achieves highly competitive performance on standard HOI detection and generation benchmarks. Specifically, UniHOI improves accuracy by 4.9% on long-tailed HOI detection and boosts interaction metrics by 42% on open-vocabulary generation tasks.

Related Work

Text-to-Image Models

Diffusion-based T2I models (Nichol et al. 2021; Saharia et al. 2022; Rombach et al. 2022; Ramesh et al. 2022) iteratively denoise latent text-conditioned embeddings (Radford et al. 2021; Raffel et al. 2020), generating high-resolution images (Rombach et al. 2022; Podell et al. 2024; Saharia et al. 2022). Recent studies (Kim et al. 2023; Mo et al. 2024; Mokady et al. 2022; Park et al. 2025) reveal that their intermediate representations encode rich semantics, supporting advanced image editing via feature-text interactions. Meanwhile, MLLM-based T2I models (Wu et al. 2024b; Yang et al. 2025; Chen et al. 2025; Xie et al. 2024) leverage large language models for enhanced multimodal understanding and unified vision-language reasoning and generation. For instance, MMaDA (Yang et al. 2025) adopts a unified diffusion framework for joint inference and generation, while Liquid (Wu et al. 2024b) shows mutual benefits between understanding and generation within MLLM architectures. Based on these works, we explore how interaction-aware cross attention in MLLMs perceives and generates interaction semantics.

HOI Understanding

HOI understanding encompasses two core tasks: HOI detection and HOI image generation. HOI detection (Ma et al. 2023; Yuan et al. 2022; Wang et al. 2022; Kim et al. 2021) seeks to localize humans and objects and classify their interactions in the form of triplets (*e.g.*, person, play, skateboard). Despite notable progress, this task remains constrained by the limited availability of high-quality annotated data. As the inverse task of detection, HOI image generation aims to generate images depicting specified interactions. Early approaches, such as InteractGAN (Gao et al. 2020), uti-

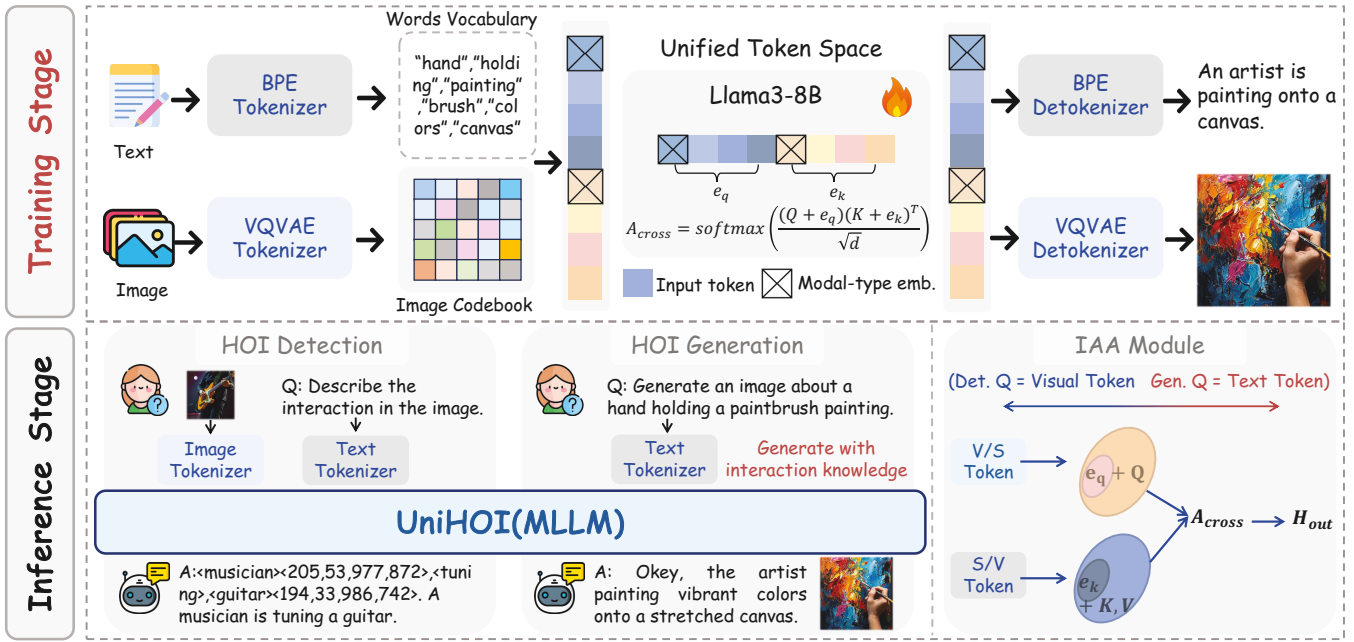


Figure 2: An overview of the UniHOI pipeline. The bottom-right shows the details of the IAA module, illustrating the bidirectional transformation between text tokens and visual tokens.

lize human poses and object references, while subsequent methods (Gao et al. 2020; Hoe et al. 2024) often depend on predefined spatial layouts alongside textual cues. However, such layout- or reference-based dependencies fundamentally constrain scalability and diversity. The requirement for expensive annotation and limited semantic expressiveness make it difficult to model complex or long-tailed interactions and to capture nuanced intentions from natural language. In this work, we aim to bridge the gap between these two tasks within a unified framework, alleviating the limitations imposed by scarce fine-grained annotations on HOI understanding, while also reducing the reliance on layout constraints for HOI image generation.

Method

Overall Framework

UniHOI is a unified MLLM that bridges HOI detection and interaction-aware image generation within a shared multimodal latent space. Built upon a sequence-to-sequence Transformer, our architecture employs a unified vocabulary that incorporates both discrete visual tokens (e.g., from VQGAN (Esser 2020)) and semantic tokens (e.g., from text captions or structured HOI triplets), enabling bidirectional vision-language modeling, shown in Figure 2.

As illustrated by the following unified formulation,

$$\mathcal{I} \rightleftharpoons \{\mathcal{H}, \mathcal{B}, \mathcal{C}\} \rightleftharpoons \mathcal{T} \quad (1)$$

where \mathcal{I} is the input image, \mathcal{H} denotes HOI triplets, \mathcal{B} the corresponding bounding boxes, \mathcal{C} the interaction caption, and \mathcal{T} a free-form or structured semantic prompt. UniHOI unifies HOI detection (image-to-triplet/caption/box)

and HOI image generation (caption-to-image) within a single model, enabling flexible and direct mapping between visual inputs and interaction semantics for mutually enhanced recognition and generation.

Modality-Aware Unified Token Space

We present a modality-aware unified token space that enables direct bidirectional transformation between visual and semantic modalities. Unlike prior multimodal alignment approaches (Liu et al. 2023; Li et al. 2023a), which confine alignment to the embedding level and only support cross-modal similarity comparison, our method unifies both modalities at the token level by constructing a shared, discrete vocabulary and a common embedding manifold.

Formally, we define the joint vocabulary as

$$\mathcal{V} = \mathcal{V}_{\text{code}} \cup \mathcal{V}_{\text{sem}}, \quad \mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d}, \quad (2)$$

where $\mathcal{V}_{\text{code}}$ and \mathcal{V}_{sem} denote the sets of visual tokens (e.g., from images) and semantic tokens (e.g., text or HOIO triplets), respectively. All tokens share a unified, trainable embedding matrix, ensuring joint representation and generation.

Crucially, for each token, we introduce a learnable **modality-type embedding** in addition to the shared token embedding, explicitly encoding modality information. The final input embedding for each token is the sum of its token embedding and modality-type embedding, allowing the Transformer to distinguish and leverage both sources effectively. This explicit type encoding helps the model disambiguate token distributions, prevents modality confusion, and enables controllable cross-modal generation, whereas methods like Liquid (Wu et al. 2024b) simply merge vocabularies without explicit modality conditioning.

With this design, both visual and semantic token sequences can serve as input or output for the same Transformer model, supporting the following explicit bidirectional mapping:

$$\underbrace{\mathbf{x}_v \in \mathcal{V}_{\text{code}}^L}_{\text{visual tokens}} \begin{array}{c} \xleftarrow{\text{detection}} \\ \xrightarrow{\text{generation}} \end{array} \underbrace{\mathbf{x}_s \in \mathcal{V}_{\text{sem}}^K}_{\text{semantic tokens}} \quad (3)$$

The unified space not only facilitates direct cross-modal generation but also enables flexible compositional reasoning within a single architecture, going beyond the capabilities of conventional token or embedding alignment frameworks.

Interaction-Aware Attention Module

To robustly capture the structured relationships underlying human-object interactions across both HOI detection (vision-to-semantics) and image generation (semantics-to-vision), we propose an Interaction-Aware Attention (IAA) module, which is designed to support both HOI detection and image generation tasks effectively. Prior approaches (Hoe et al. 2024; Li et al. 2023b) typically employ asymmetric architectures or separate attention modules for detection and generation, which restrict flexibility. In contrast, our IAA module adopts a parameter-shared, symmetric cross-attention mechanism that supports both tasks within a single structure, which enables consistent, context-aware alignment between vision and semantics.

Design of Symmetric Cross-Attention. IAA is implemented as a single, parameter-shared cross-attention block, whose directionality is determined solely by the task: in detection, *visual tokens* acts as queries and *semantic tokens* serves as keys and values; for generation, this assignment is reversed. Each token is enriched with a learnable *modality-type embedding* \mathbf{e}_t , indicating whether it originates from the visual or semantic space, helping the model distinguish between heterogeneous sources during attention calculation.

Formally, let \mathbf{Q} , \mathbf{K} , and \mathbf{V} be the query, key, and value matrices (the assignment depends on the task direction), and let \mathbf{e}_q , \mathbf{e}_k be modality-type embeddings for queries and keys. The cross-modal attention in IAA is defined as:

$$\mathbf{A}_{\text{cross}} = \text{softmax} \left(\frac{(\mathbf{Q} + \mathbf{e}_q)(\mathbf{K} + \mathbf{e}_k)^T}{\sqrt{d}} \right) \mathbf{V} \quad (4)$$

$$\mathbf{H}_{\text{out}} = \mathbf{A}_{\text{cross}} + \mathbf{Q} \quad (5)$$

Here, Eq. (4) computes attention with awareness of each token’s modality, and Eq. (5) applies a residual connection to preserve original query information.

Adaptability and Effectiveness. Although we use only coarse-grained modality-type embeddings (“visual” or “semantic”), the IAA module exploits the structured HOI triplet input and the capacity of self-attention to implicitly align semantic slots with visual regions. Notably, our strictly symmetric design shares cross-attention architecture and parameters for both HOI detection and image generation, supporting flexible input-role swapping. As shown in Figure 3, IAA accurately captures the spatial distribution of HOI triplets across both inverse tasks, enabling efficient and fine-grained cross-modal reasoning.

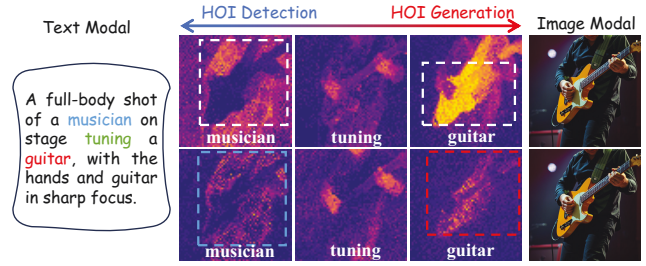


Figure 3: Visualization of interaction-aware attention maps produced by IAA for HOI detection and generation tasks. Bidirectional arrows indicate the mutual mapping between visual and textual tokens, highlighting IAA’s capability in cross-modal interactive semantic modeling. The transitions among prompts, images, and HOI triplets further demonstrate unified token transformations across inverse tasks.

Semi-supervised Learning Strategy

Unified Cycle Consistency in Token Space. We leverage the unified token space to enable effective learning from fully, weakly, or unlabeled data, and to support both HOI detection and interaction-aware image generation as mutually inverse tasks. Unlike prior methods that rely on pseudo-labeling or disjoint modality-specific objectives, our approach introduces a dual cycle-consistency framework that enables joint training with heterogeneous supervision.

Specifically, let \mathcal{F} and \mathcal{G} denote the detection and generation modules, respectively; $\mathcal{T} \in \mathcal{V}_{\text{sem}}$ denotes semantic token sequences (e.g., HOI triplets), and $\mathcal{I} \in \mathcal{V}_{\text{code}}$ denotes visual tokens. We formulate the cross-modal cycle losses as:

$$\begin{aligned} \mathcal{L}_{\text{Cycle}}^{\text{sem} \rightarrow \text{code} \rightarrow \text{sem}} &= \mathbb{E}_{\mathcal{T}} [\|\mathcal{F}(\mathcal{G}(\mathcal{T})) - \mathcal{T}\|_1] \\ \mathcal{L}_{\text{Cycle}}^{\text{code} \rightarrow \text{sem} \rightarrow \text{code}} &= \mathbb{E}_{\mathcal{I}} [\|\mathcal{G}(\mathcal{F}(\mathcal{I})) - \mathcal{I}\|_1] \end{aligned} \quad (6)$$

This dual cycle-consistency objective enforces that both semantic-to-visual-to-semantic and visual-to-semantic-to-visual mappings remain consistent in the unified token space. For each sample, the model either reconstructs semantics from generated visual tokens or reconstructs visuals from inferred semantic tokens. The cycle loss measures the difference between the initial and reconstructed token sequences in both directions. This unified training enables efficient use of data with any level of supervision, facilitating robust and effective learning of detection and generation tasks even with limited annotations.

Model Training

UniHOI is trained end-to-end with a semi-supervised paradigm that fully exploits the intrinsic duality between HOI detection and generation within a unified token space. Instead of optimizing detection and generative branches separately, all supervision—regardless of annotation granularity—is formulated as mutually inverse, cycle-consistent transformations and reconstructions in this unified space, enabling maximal alignment between original and reconstructed modalities.

Formally, for any input x (image, HOI triplet, or interaction caption), we define the unified cycle mapping:

$$x \xrightarrow{\mathcal{F}} y \xrightarrow{\mathcal{G}} \hat{x}, \quad \mathcal{L}_{\text{cycle}} = d(x, \hat{x}), \quad (7)$$

where \mathcal{F} and \mathcal{G} are invertible, parameter-shared detection and generation functions, $d(\cdot)$ denotes the sequence-level cross-entropy loss between the original and reconstructed tokens, with all tokens residing in the same unified space.

This framework naturally accommodates heterogeneous data: for paired data, both directions are optimized to enforce alignment; for single-modality or unlabeled data, available information is propagated through the cycle for reconstruction, obviating the need for handcrafted pseudo-labels or task-specific losses.

Unified Loss Formulation. The overall loss aggregates all cycle pathways with dynamic weights, together with semantic alignment and diversity objectives:

$$\mathcal{L}_{\text{uni}} = \sum_{i=1}^N \lambda^{(i)} \mathcal{L}_{\text{cycle}}^{(i)} + \mathcal{L}_{\text{sem_align}} + \mathcal{L}_{\text{div}}, \quad (8)$$

where N is the number of cycle pathways (e.g., for paired, unpaired, or weakly supervised samples), $\mathcal{L}_{\text{cycle}}^{(i)}$ is the cycle-consistency loss for the i -th data type, and the weighting coefficients $\lambda^{(i)}$ are set to $\lambda^{(i)} = 1$. $\mathcal{L}_{\text{sem_align}}$ encourages semantic and spatial consistency, while \mathcal{L}_{div} promotes generation diversity. Both are computed from attention maps to leverage spatial and statistical properties during training.

Experiments

Datasets

UniHOI is trained semi-supervisedly using a curated mixture of three data types to balance supervision and scalability: **10%** strongly-supervised, **25%** weakly-supervised, and **65%** unsupervised data. Compared to HICO-DET, our dataset not only covers nearly all its interaction types, but also contains a substantially larger and more diverse set of HOI triplets mined from open-domain sources, which facilitates truly open-world HOI understanding.

Strongly-supervised: We use HICO-DET (Chao et al. 2017) and V-COCO (Hou et al. 2020), which provide detailed HOI triplet annotations with bounding boxes and role labels, supplying rich spatial-semantic supervision.

Weakly-supervised: 150K image-text pairs from LAION-SG (Li et al. 2024) are used, where triplets extracted from scene graphs serve as weak HOI signals, alleviating the long-tail annotation problem.

Unsupervised: 400K image-text pairs from LAION-400M (Schuhmann et al. 2021) are employed, utilizing only the captions for contrastive vision-language learning, without any structured labels.

Evaluation Metrics

HOI Detection: We report mean Average Precision (mAP) on HICO-DET (Chao et al. 2017) and V-COCO (Hou et al. 2020), focusing on interaction instance mAP under both Default and Rare settings.

HOI Generation: We assess generation quality using 10K prompts from our test set to compute Wise Score (Niu et al. 2025), Image Reward (Xu et al. 2023), FID, and CLIP Score (Radford et al. 2021). For HOI-specific evaluation, we report **HOI Score**, which measures the accuracy of the detected triplet against the input triplet using a pretrained HOI detector (Ma et al. 2023), and **Interaction Accuracy**, which assesses how well the generated interaction details match the prompt by comparing a caption derived from the detector’s output with the input prompt. Additionally, we include GenEval (Lin et al. 2024) for comprehensive assessment.

Implementation Details

In our experiments, we employ Llama3-8B (Grattafiori et al. 2024) as the backbone of UniHOI, and adopt VQ-GAN (Esser 2020) from Chameleon (Team 2024) as the image tokenizer. Model finetuning is performed using the Adam optimizer with a constant learning rate of 5×10^{-4} , following a linear warm-up over the first 10,000 iterations. UniHOI is trained for 700,000 iterations in total, with a per-device batch size of 8 on 32 NVIDIA H800 GPUs; to further increase the effective batch size to 16, we employ gradient accumulation with a step size of 2. A temperature parameter of $\tau = 0.07$ is consistently used across all experiments. To mitigate the potential inefficiency arising from semi-supervised learning, we utilize a hybrid data loading strategy that ensures balanced sampling from both labeled and unlabeled datasets within each mini-batch.

Quantitative & Qualitative Results

Quantitative Results on HOI Detection. Table 1 presents a comprehensive comparison with state-of-the-art methods on HICO-DET and V-COCO benchmarks. UniHOI establishes new state-of-the-art results on HICO-DET, achieving 48.16 mAP on the full test set, 50.74 mAP on rare categories, and 51.34 mAP under the Known Object setting. On V-COCO, UniHOI achieves 72.91 in AP_{role}^{S1} and the highest AP_{role}^{S2} score of 77.45, demonstrating clear superiority in precise role-object localization despite a slightly lower AP_{role}^{S1} compared to HOI-IDiff (Hui et al. 2025). These consistent improvements, particularly on rare and long-tailed HOI categories, further demonstrating the mutual benefits of jointly modeling HOI detection and generation.

Quantitative Results on HOI Generation. Table 2 presents a comprehensive quantitative comparison on HOI-oriented image generation benchmarks. UniHOI consistently achieves superior results across all major metrics, including the best Image Reward (1.17), lowest FID score (18.2), and highest CLIP Score (32.46), highlighting its strong perceptual quality and text-image correspondence. UniHOI also establishes new state-of-the-art performance on all GenAIEval submetrics, demonstrating excellent capability in compositional and spatial understanding. In terms of interaction-focused metrics such as HOI Score and Interaction Accuracy, our method significantly outperforms existing approaches, surpassing the previous state-of-the-art by 14.3% and 42.0%, respectively. This highlights the strong capability of our approach in generating images with fine-

Method	Backbone	HICO-DET						V-COCO	
		Default			Known Object			AP_{role}^{S1}	AP_{role}^{S2}
		Full	Rare	Non-rare	Full	Rare	Non-rare		
BCOM (Wang et al. 2024)	R50+CLIP	39.34	39.90	39.17	42.24	42.86	42.05	65.10	69.90
MP-HOI (Yang et al. 2024)	Swin-L	44.53	44.48	44.55	-	-	-	66.22	67.64
SICHOI (Luo et al. 2024)	R101+ViT-L/16	45.04	45.61	44.88	48.16	48.37	48.09	71.13	75.62
PA-HOI (Wu et al. 2024a)	Swin-L	46.01	46.74	45.80	49.50	50.59	49.18	63.04	68.75
HOI-IDiff (Hui et al. 2025)	Diffusion	47.71	48.36	47.52	50.56	51.95	50.14	73.42	76.13
UniHOI (Ours)	VQGAN	48.16	50.74	48.12	51.34	53.72	50.33	72.91	77.45

Table 1: Comparison of state-of-the-art methods on HICO-DET (Default / Known Object) and V-COCO benchmarks (mAP scores). Best results are highlighted in bold. AP_{role}^{S1} and AP_{role}^{S2} denote standard splits on V-COCO.

Method	Wise Score \uparrow	Image Reward \uparrow	FID Score \downarrow	CLIP Score \uparrow	HOI Score \uparrow	Interaction Accuracy \uparrow	GenAIEval \uparrow		
							Single Obj.	Two Obj.	Position
∇ Generation Only									
InteractDiffusion (Hoe et al. 2024)	-	0.79	38.2	13.43	0.40	0.22	0.71	0.34	0.07
DALL-E2 (Ramesh et al. 2022)	-	0.83	28.6	25.20	0.48	0.29	0.94	0.66	0.10
SDXL (Podell et al. 2024)	0.43	1.13	19.1	30.87	0.54	0.38	0.98	0.74	0.15
SDv3.5 (Esser et al. 2024)	0.51	1.15	17.7	31.54	0.56	0.35	0.96	0.71	0.14
∇ Unified Understanding & Generation									
Chameleon (Team 2024)	-	0.83	27.3	20.32	0.41	0.28	-	-	-
Show-o (Xie et al. 2024)	0.28	0.92	24.7	28.94	0.46	0.31	0.95	0.52	0.11
Janus (Wu et al. 2025)	0.16	1.03	22.1	29.45	0.50	0.36	0.97	0.68	0.28
Liquid (Wu et al. 2024b)	-	-	25.8	21.73	0.39	0.26	-	-	-
VAR-GPT (Zhuang et al. 2025)	-	0.94	23.8	28.85	0.44	0.33	0.96	0.53	0.13
UniHOI (Ours)	0.50	1.17	18.2	32.46	0.64	0.54	0.99	0.76	0.42

Table 2: Evaluation results on HOI-oriented image generation benchmarks. We report Wise Score, Image Reward, FID, and CLIP Score to assess perceptual quality and text-image correspondence; HOI Score and Interaction Accuracy (IA) for the correctness of human-object interactions; and GenAIEval submetrics (SingleObj, TwoObj, Position) for compositional and spatial understanding. Lower FID and higher values for all other metrics indicate better performance.

grained interactions. These results collectively validate the effectiveness of our unified framework in generating high-quality images that accurately represent complex human-object interactions.

Qualitative Results. Shown in Figure 4, compared to state-of-the-art models in HOI detection, our method demonstrates superior fine-grained perception when handling hard cases. By leveraging open-world knowledge acquired under a unified label space, our approach extracts more accurate HOI triplets. For HOI generation, our method exhibits clear advantages in generating detailed and natural interactions compared to existing models—for instance, enabling more natural tool usage, such as picking up a paintbrush, with highly faithful hand pose information.

Ablation Study

Ablation on Unified Token Space. As shown in Table 3, models without modality-type embedding or with separated embeddings achieve consistently lower performance in both detection (Full mAP: 48.16 \rightarrow 47.62/47.03) and generation (HOI Score: 0.64 \rightarrow 0.57/0.44), verifying that our unified, modality-aware token space substantially enhances both tasks by enabling more effective information fusion across modalities.

Ablation on Interaction-Aware Attention. As presented

Method	HICO-DET (Default) \uparrow		Generation \uparrow	
	Full	Rare	HOI Score	IA
Separate Emb.	47.03	48.38	0.44	0.32
Shared Emb. Only	47.48	48.92	0.52	0.41
w/o Type Emb.	47.62	49.13	0.57	0.46
Ours (Full)	48.16	50.74	0.64	0.54

Table 3: Ablation on **unified token space**. “Separate Emb.” uses unshared visual and text embeddings; “Shared Emb. Only” shares embeddings but not vocabularies; “w/o Type Emb.” unifies vocab and embeddings but removes type information;

in Table 4, substituting IAA with a standard cross-attention module or removing modality-type embedding consistently degrades performance across all metrics, especially on rare HOIs and interaction accuracy (e.g., IA: 0.54 \rightarrow 0.49/0.39). This demonstrates that our symmetric, modality-aware attention design is crucial for effective and generalizable cross-modal modeling. More ablation studies are provided in the supplementary material.

Ablation on Supervision Ratio. We further examine the cross-task benefits between HOI detection and generation under varying supervision ratios. Table 5 demonstrates that



Figure 4: Qualitative results of UniHOI. For HOI detection, UniHOI demonstrates enhanced fine-grained interaction understanding; for HOI generation, it produces detailed interactive scenes, including realistic hand poses and precise tool usage.

Setting	HICO-DET (Default) \uparrow		Generation \uparrow	
	Full	Rare	HOI Score	IA
w/o IAA	47.10	49.03	0.51	0.39
w/o Type Emb.	47.71	49.55	0.59	0.49
Ours(Full)	48.16	50.74	0.64	0.54

Table 4: Ablation on **interaction-aware attention**. “w/o IAA” denotes removing interaction-aware attention; “w/o Type Emb.” removes modality-type embedding from the symmetric module. Generation columns report HOI Score and Interaction Accuracy.

increasing the proportions of weakly and unsupervised data boosts performance on both HOI detection (HOI Det.) and HOI generation (HOI Gen.). These results highlight a mutual promotion between detection and generation tasks, where supervision from one task benefits the other, proving the effectiveness of our multi-task and semi-supervised learning strategy.

Conclusion

In this paper, we present UniHOI, a unified semi-supervised multimodal framework that jointly addresses human-object interaction (HOI) detection and generation via a modality-aware token space and symmetric interaction-aware attention modules. By unifying visual and semantic representa-

Strong	Weak	Unsupervised	HOI Det.	HOI Gen.
100%	–	–	45.62	0.22
50%	20%	30%	46.43	0.38
20%	20%	60%	47.65	0.59
10%	25%	65%	48.16	0.64

Table 5: Ablation on **supervision ratios**. Varying the proportion of strong, weak, and unsupervised data shows the effect on HOI detection and generation. HOI Det. denotes the Full metric under the Default category in HICO-DET, and HOI Gen. denotes the HOI Score.

tions at the token level and leveraging a symmetric cross-modal attention mechanism, UniHOI achieves robust and generalizable performance for both tasks, substantially advancing the state of the art on benchmark datasets. Our approach enables effective knowledge and data sharing across HOI detection and generation, thereby improving data efficiency and long-tail generalization in open-world scenarios. Extensive experiments demonstrate that unified tokenization is crucial for flexible, compositional, and accurate cross-modal reasoning. We believe UniHOI not only advances the unified modeling of inverse HOI tasks, but also provides new insights for bridging recognition and generation in broader multimodal and open-vocabulary contexts.

Acknowledgments

This work was supported by the STI 2030-Major Projects under Grant No. 2022ZD0208801 and the NSFC under grant No. 62088102.

References

- Chao, Y.-W.; Liu, Y.; Liu, M. X.; Zeng, H.; and Deng, J. 2017. Learning to Detect Human-Object Interactions. In *WACV*.
- Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; and Ruan, C. 2025. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. *arXiv:2501.17811*.
- Esser, P. 2020. Taming Transformers for High-Resolution Image Synthesis. *arXiv:2012.09841*.
- Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.
- Gao, C.; Liu, S.; Zhu, D.; LIU, Q.; Cao, J.; He, H.; He, R.; and Yan, S. 2020. InteractGAN: Learning to Generate Human-Object Interaction. In *ACM MM*.
- Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Hoe, J. T.; Jiang, X.; Chan, C. S.; Tan, Y.-P.; and Hu, W. 2024. InteractDiffusion: Interaction Control in Text-to-Image Diffusion Models. In *CVPR*.
- Hou, Z.; Peng, X.; Qiao, Y.; and Tao, D. 2020. Visual Compositional Learning for Human-Object Interaction Detection. In *ECCV*.
- Hui, X.; Qu, H.; Rahmani, H.; and Liu, J. 2025. An Image-like Diffusion Method for Human-Object Interaction Detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14002–14012.
- Kim, B.; Lee, J.; Kang, J.; Kim, E.-S.; and Kim, H. J. 2021. HOTR: End-to-End Human-Object Interaction Detection with Transformers.
- Kim, Y.; Lee, J.; Kim, J.-H.; Ha, J.-W.; and Zhu, J.-Y. 2023. Dense Text-to-Image Generation with Attention Modulation. In *ICCV*.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023b. GLIGEN: Open-Set Grounded Text-to-Image Generation. In *CVPR*.
- Li, Z.; Meng, C.; Li, Y.; Yang, L.; Zhang, S.; Ma, J.; Li, J.; Yang, G.; Yang, C.; Yang, Z.; Chang, J.; and Sun, L. 2024. LAION-SG: An Enhanced Large-Scale Dataset for Training Complex Image-Text Models with Structural Annotations. *arXiv:2412.08580*.
- Lin, Z.; Pathak, D.; Li, B.; Li, J.; Xia, X.; Neubig, G.; Zhang, P.; and Ramanan, D. 2024. Evaluating Text-to-Visual Generation with Image-to-Text Generation. *arXiv preprint arXiv:2404.01291*.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning.
- Luo, J.; Ren, W.; Jiang, W.; Chen, X.; Wang, Q.; Han, Z.; and Liu, H. 2024. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28212–28222.
- Ma, S.; Wang, Y.; Wang, S.; and Wei, Y. 2023. FGAHOI: Fine-Grained Anchors for Human-Object Interaction Detection. *IEEE TPAMI*, 2415–2429.
- Mo, S.; Mu, F.; Lin, K. H.; Liu, Y.; Guan, B.; Li, Y.; and Zhou, B. 2024. FreeControl: Training-Free Spatial Control of Any Text-to-Image Diffusion Model with Any Condition. In *CVPR*.
- Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Null-text Inversion for Editing Real Images using Guided Diffusion Models. In *arXiv preprint arXiv:2211.09794*.
- Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*.
- Niu, Y.; Ning, M.; Zheng, M.; Jin, W.; Lin, B.; Jin, P.; Liao, J.; Feng, C.; Ning, K.; Zhu, B.; et al. 2025. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*.
- Park, J.; Ko, J.; Byun, D.; Suh, J.; and Rhee, W. 2025. Cross-Attention Head Position Patterns Can Align with Human Visual Concepts in Text-to-Image Generative Models. In *ICLR*.
- Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Muller, J.; Penna, J.; and Rombach, R. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *ICLR*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.
- Raffel, C.; Shazeer, N. M.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*.
- Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. In *arXiv preprint arxiv:2204.06125*.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*.
- Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *NeurIPS*.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *arXiv preprint arxiv:2111.02114*.

Team, C. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Wang, G.; Guo, Y.; Xu, Z.; and Kankanhalli, M. 2024. Bilateral adaptation for human-object interaction detection with occlusion-robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27970–27980.

Wang, S.; Duan, Y.; Ding, H.; Tan, Y.-P.; Yap, K.-H.; and Yuan, J. 2022. Learning Transferable Human-Object Interaction Detectors with Natural Language Supervision. In *CVPR*.

Wu, C.; Chen, X.; Wu, Z.; Ma, Y.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; Ruan, C.; et al. 2025. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12966–12977.

Wu, E. Z.; Li, Y.; Wang, Y.; and Wang, S. 2024a. Exploring Pose-Aware Human-Object Interaction via Hybrid Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17815–17825.

Wu, J.; Jiang, Y.; Ma, C.; Liu, Y.; Zhao, H.; Yuan, Z.; Bai, S.; and Bai, X. 2024b. Liquid: Language models are scalable multi-modal generators. *arXiv preprint arXiv:2412.04332*.

Xie, J.; Mao, W.; Bai, Z.; Zhang, D. J.; Wang, W.; Lin, K. Q.; Gu, Y.; Chen, Z.; Yang, Z.; and Shou, M. Z. 2024. Show-o: One Single Transformer to Unify Multimodal Understanding and Generation. *arXiv preprint arXiv:2408.12528*.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 15903–15935.

Yang, J.; Li, B.; Zeng, A.; Zhang, L.; and Zhang, R. 2024. Open-world human-object interaction detection via multi-modal prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16954–16964.

Yang, L.; Tian, Y.; Li, B.; Zhang, X.; Shen, K.; Tong, Y.; and Wang, M. 2025. MMaDA: Multimodal Large Diffusion Language Models. *arXiv:2505.15809*.

Yuan, H.; Jiang, J.; Albanie, S.; Feng, T.; Huang, Z.; Ni, D.; and Tang, M. 2022. RLIP: Relational Language-Image Pre-training for Human-Object Interaction Detection. In *NeurIPS*.

Zhuang, X.; Xie, Y.; Deng, Y.; Liang, L.; Ru, J.; Yin, Y.; and Zou, Y. 2025. VARGPT: Unified Understanding and Generation in a Visual Autoregressive Multimodal Large Language Model. *arXiv preprint arXiv:2501.12327*.