Novelty Assessment Report

Paper: TOUCH: Text-guided Controllable Generation of Free-Form Hand-Object Interactions

PDF URL: https://openreview.net/pdf?id=4VW9HVCRw0

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-30

Abstract

Hand-object interaction (HOI) is fundamental for humans to express intent. Existing HOI generation research is predominantly confined to fixed grasping patterns, where control is tied to physical priors such as force closure or generic intent instructions, even when expressed through elaborate language. Such an overly general conditioning imposes a strong inductive bias for stable grasps, thus failing to capture the diversity of daily HOI. To address these limitations, we introduce $\textbf{Free-Form HOI Generation}$, which aims to generate controllable, diverse, and physically plausible HOI conditioned on fine-grained intent, extending HOI from grasping to free-form interactions, like pushing, poking, and rotating. To support this task, we construct $\textbf{WildO2}$, an in-the-wild diverse 3D HOI dataset, which includes diverse HOI derived from internet videos. Specifically, it contains 4.4k unique interactions across 92 intents and 403 object categories, each with detailed semantic annotations. Building on this dataset, we propose $\textbf{TOUCH}$, a three-stage framework centered on a multi-level diffusion model that facilitates fine-grained semantic control to generate versatile hand poses beyond grasping priors. This process leverages explicit contact modeling for conditioning and is subsequently refined with contact consistency and physical constraints to ensure realism. Comprehensive experiments demonstrate our method's ability to generate controllable, diverse, and physically plausible hand interactions representative of daily activities.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Text-Guided Controllable Generation of Free-Form Hand-Object Interactions**

A total of **39 papers** were analyzed and organized into a taxonomy with **21 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Interaction Synthesis Approaches**
- **Interaction Representation and Decomposition**
- **Specialized Interaction Contexts**
- **Data Collection and Annotation**
- **Affordance and Contact Modeling**
- **Robotic Manipulation Applications**
- **Related Motion Generation Tasks**

Complete Taxonomy Tree

- Text-Guided Controllable Generation of Free-Form Hand-Object Interactions Survey Taxonomy
- Interaction Synthesis Approaches
  - Diffusion-Based Interaction Generation
  - Dual-Branch and Modular Diffusion Architectures (3 papers)
    - [6] Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models (Xiaogang Peng, 2025) View paper
    - [10] Cg-hoi: Contact-guided 3d human-object interaction generation (Christian Diller, 2024) View paper
    - [12] Thor: Text to human-object interaction diffusion via relation intervention (Qianyang Wu, 2024) View paper
  - Contact-Guided and Constraint-Based Diffusion ★ (3 papers)
    - [0] TOUCH: Text-guided Controllable Generation of Free-Form Hand-Object Interactions (Anon et al., 2026) View paper
    - [9] Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions (Sammy Christen, 2024) View paper
    - [18] HOIDiNi: Human-Object Interaction through Diffusion Noise Optimization (Tevet, 2025) View paper
  - Staged and Temporal Diffusion Processes (2 papers)
    - [1] Controllable human-object interaction synthesis (Li, 2024) View paper
    - [31] OASIS: Object-guided Attention for Text-conditional Diffusion Synthesis of Human Interaction Sequences (Chih-Chun Yang, 2025) View paper
  - Multimodal and Image-Conditioned Diffusion (3 papers)
    - [3] Hoidiffusion: Generating realistic 3d hand-object interaction data (Mengqi Zhang, 2024) View paper
    - [26] MEgoHand: Multimodal Egocentric Hand-Object Interaction Motion Generation (Zhou, 2025) View paper
    - [32] InterFusion: Text-Driven Generation of 3D Human-Object Interaction (Dai Sisi, 2024) View paper
  - Joint-Level and Kinematic Chain Modeling (2 papers)
  - [5] Chainhoi: Joint-based kinematic chain modeling for human-object interaction generation (Ling-An Zeng, 2025) View paper
  - [17] Efficient Explicit Joint-level Interaction Modeling with Mamba for Text-guided HOI Generation (Guohong Huang, 2025) View paper
  - LLM and Token-Based Generation (2 papers)
  - [16] HOIGPT: Learning Long-Sequence Hand-Object Interaction with Language Models (Mingzhen Huang, 2025) View paper

  - [23] OpenHOI: Open-World Hand-Object Interaction Synthesis with Multimodal Large Language Model (Zhang Zhen-Hao, 2025) View paper
  - Zero-Shot and Out-of-Domain Generalization (3 papers)
  - [7] Interdreamer: Zero-shot text to 3d dynamic human-object interaction (Xu, 2024) View paper
  - [13] OOD-HOI: Text-driven 3d whole-body human-object interactions generation beyond training domains (Zhang Yi-xuan, 2024) View paper
  - [14] InteractAnything: Zero-shot Human Object Interaction Synthesis via LLM Feedback and Object Affordance Parsing (Jinlu Zhang, 2025) View paper
- Interaction Representation and Decomposition
  - Subtask Decomposition and Two-Stage Frameworks (3 papers)
  - [4] SIGHT: Single-Image Conditioned Generation of Hand Trajectories for Hand-Object Interaction (Gavryushin, 2025) View paper
  - [24] Text2HOI: Text-Guided 3D Motion Generation for Hand-Object Interaction (Junuk Cha, 2024) View paper
  - [34] SIGHT: Synthesizing Image-Text Conditioned and Geometry-Guided 3D Hand-Object Trajectories (Gavryushin, 2025) View paper
  - Action Basis and Invariant Reasoning (1 papers)
  - [35] EigenActor: Variant Body-Object Interaction Generation Evolved from Invariant Action Basis Reasoning (Gao Xuehao, 2025) View paper
  - Part-Level Semantic Alignment (1 papers)
  - [30] Textim: Part-aware interactive motion synthesis from text (Fan Siyuan, 2024) View paper
- Specialized Interaction Contexts
  - Scene-Aware and Contextual Interaction (3 papers)
  - [11] Human-object interaction from human-level instructions (Wu Zhen, 2025) View paper
  - [20] Generating human interaction motions in scenes with text control (Yi, 2024) View paper
  - [33] InteractMove: Text-Controlled Human-Object Interaction Generation in 3D Scenes with Movable Objects (Xinhao Cai, 2025) View paper
  - Dual-Hand and Bimanual Grasping (1 papers)
  - [37] DHAGrasp: Synthesizing Affordance-Aware Dual-Hand Grasps with Text Instructions (Li, 2025) View paper
  - Functional and Semantic Grasp Generation (2 papers)
  - [36] FunHOI: Annotation-Free 3D Hand-Object Interaction Generation via Functional Text Guidanc (Tian Yong-qi, 2025) View paper
  - Multiple Object and Long-Horizon Interaction (2 papers)
  - [15] Hoianimator: Generating text-prompt human-object animations using novel perceptive diffusion models (Wenfeng Song, 2024) View paper
  - [29] Himo: A new benchmark for full-body human interacting with multiple objects (Lv, 2024) View paper
- Data Collection and Annotation
  - Motion Capture and Multi-View Systems (1 papers)
  - [19] Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions (Jeonghwan Kim, 2025) View paper
  - Video-Based Dataset Construction (1 papers)
  - [2] TASTE-Rob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation (Hongxiang Zhao, 2025) View paper
  - Synthetic Data Generation and Augmentation (1 papers)
  - [39] Prompt-Propose-Verify: A Reliable Hand-Object-Interaction Data Generation Framework using Foundational Models (Gurusha Juneja, 2023) View paper
- Affordance and Contact Modeling
  - Text-Driven Affordance Learning (1 papers)
  - [8] Text-driven affordance learning from egocentric vision (Tomoya Yoshida, 2025) View paper
  - Natural Language Guided Contact Modeling (2 papers)
  - [27] TIGeR: Text-Instructed Generation and Refinement for Template-Free Hand-Object Interaction (Zheng, 2025) View paper
  - [28] Nl2contact: Natural language guided 3d hand-object contact modeling with diffusion model (Zhang Zhongqun, 2024) View paper
- Robotic Manipulation Applications (2 papers)
  - [21] Multimodal human-intent modeling for contextual robot-to-human handovers of arbitrary objects (Lucas Chen, 2025) View paper
  - [22] Ram: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation (Yuxuan Kuang, 2024) View paper
- Related Motion Generation Tasks (1 papers)
  - [25] Prompting future driven diffusion model for hand motion prediction (Bowen Tang, 2024) View paper

Narrative

Core task: text-guided controllable generation of free-form hand-object interactions. The field has organized itself around several complementary branches that address different facets of synthesizing realistic hand-object interactions from natural language descriptions. Interaction Synthesis Approaches encompasses the algorithmic strategies—ranging from diffusion-based methods like Hoidiffusion[3] and Diffh2o[9] to transformer and state-space architectures such as Mamba HOI[17]—that generate plausible motion sequences. Interaction Representation and Decomposition focuses on how to encode and structure the problem, often breaking interactions into contact patterns, grasp phases, or temporal stages as seen in works like Chainhoi[5]. Specialized Interaction Contexts targets domain-specific scenarios such as robotic handovers or egocentric manipulation, while Data Collection and Annotation and Affordance and Contact Modeling provide the foundational resources and geometric reasoning needed to ground these generations in physical plausibility. Robotic Manipulation Applications bridges the gap to real-world deployment, and Related Motion Generation Tasks situates this work within the broader landscape of human motion synthesis.

Within the diffusion-based synthesis branch, a particularly active line of research emphasizes contact-guided and constraint-based generation to ensure physical realism and fine-grained control. TOUCH[0] exemplifies this direction by incorporating explicit contact constraints into the diffusion process, enabling more precise manipulation of where and how the hand engages with objects. This approach contrasts with earlier diffusion methods like Hoidiffusion[3], which may rely more heavily on learned priors without explicit geometric guidance, and complements recent efforts such as HOIDiNi[18] that explore alternative constraint formulations. The trade-off centers on balancing generative flexibility with physical fidelity: purely data-driven diffusion can produce diverse outputs but may struggle with rare or geometrically intricate interactions, whereas contact-aware methods like TOUCH[0] sacrifice some variability to maintain tighter adherence to physical plausibility. Open questions remain around scalability to complex multi-object scenes and the integration of higher-level semantic reasoning from text, as explored in works like HOIGPT[16] and Text2HOI[24].

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions

Authors: Sammy Christen, Shreyas Hampali, S. Christen, Fadime Sener, Edoardo Remelli, et al. (10 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

We introduce DiffH2O, a new diffusion-based framework for synthesizing realistic, dexterous hand-object interactions from natural language. Our model employs a temporal two-stage diffusion process, dividing hand-object motion generation into grasping and interaction stages to enhance generalization to various object shapes and textual prompts. To improve generalization to unseen objects and increase output controllability, we propose grasp guidance, which directs the diffusion model towards a ta...

Relationship Analysis

Both papers belong to the Contact-Guided and Constraint-Based Diffusion category, employing diffusion models with explicit contact modeling for hand-object interaction generation. They overlap in using contact maps as conditioning signals and incorporating physical constraints to ensure realistic interactions. However, TOUCH focuses on free-form interactions beyond grasping with fine-grained text control via multi-level diffusion and introduces the WildO2 dataset from internet videos, while DiffH2O emphasizes a temporal two-stage diffusion process (grasping then interaction) with grasp guidance for controllability and provides detailed textual annotations for the GRAB dataset.

2. HOIDiNi: Human-Object Interaction through Diffusion Noise Optimization

Authors: Tevet, Guy, Roey Ron, Sawdayee, Haim, et al. (10 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

We present HOIDiNi, a text-driven diffusion framework for synthesizing realistic and plausible human-object interaction (HOI). HOI generation is extremely challenging since it induces strict contact accuracies alongside a diverse motion manifold. While current literature trades off between realism and physical correctness, HOIDiNi optimizes directly in the noise space of a pretrained diffusion model using Diffusion Noise Optimization (DNO), achieving both. This is made feasible thanks to our obs...

Relationship Analysis

Both papers belong to the Contact-Guided and Constraint-Based Diffusion category, employing diffusion models with explicit contact modeling and physical constraints for hand-object interaction generation. They overlap in using contact maps as conditioning signals and incorporating penetration/physical plausibility losses to ensure realistic interactions. However, TOUCH focuses on free-form interactions beyond grasping with multi-level text control and a three-stage pipeline (contact prediction, multi-level diffusion, refinement), while HOIDiNi uses a two-phase diffusion noise optimization strategy (object-centric then human-centric) with explicit contact pair prediction between hand anchors and object surfaces, optimizing directly in the noise space of a pretrained diffusion model.

Contributions Analysis

Overall novelty summary. The paper introduces free-form hand-object interaction generation, extending beyond fixed grasping patterns to diverse manipulations like pushing and poking. It resides in the Contact-Guided and Constraint-Based Diffusion leaf, which contains three papers including TOUCH itself. This leaf sits within the broader Diffusion-Based Interaction Generation branch, indicating a moderately populated research direction focused on incorporating explicit physical constraints into diffusion models. The taxonomy reveals this is an active but not overcrowded area, with sibling leaves exploring dual-branch architectures and temporal decomposition strategies.

The taxonomy structure shows TOUCH's leaf neighbors include Dual-Branch and Modular Diffusion Architectures and Staged and Temporal Diffusion Processes, both addressing complementary aspects of interaction synthesis. The broader Interaction Synthesis Approaches branch encompasses alternative paradigms like LLM-based token generation and joint-level kinematic modeling. The WildO2 dataset contribution connects to the Data Collection and Annotation branch, specifically Video-Based Dataset Construction, which contains only one other paper. This positioning suggests the work bridges generative modeling innovations with data infrastructure needs in a relatively underexplored intersection.

Among thirty candidates examined, none clearly refute the three core contributions. The free-form interaction task formulation examined ten candidates with zero refutations, suggesting novelty in extending beyond stability-focused grasping. The WildO2 dataset construction pipeline similarly showed no overlapping prior work among ten candidates, though the limited search scope means comprehensive video-based HOI datasets may exist outside this sample. The TOUCH framework's multi-level diffusion architecture examined ten candidates without refutation, indicating the specific combination of contact guidance and fine-grained semantic control appears distinctive within the examined literature.

Based on the top-thirty semantic matches and taxonomy structure, the work appears to occupy a relatively sparse intersection of contact-aware diffusion and diverse interaction modeling. The analysis covers diffusion-based synthesis methods and related dataset construction efforts but does not exhaustively survey all video-based HOI datasets or alternative generative paradigms. The absence of refutations across contributions suggests meaningful novelty within the examined scope, though the limited candidate pool precludes definitive claims about the broader literature landscape.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Free-form hand-object interaction generation task

Description: The authors propose a new task that extends hand-object interaction generation beyond traditional grasp-centric approaches to encompass diverse non-grasping manipulations such as pushing, poking, and rotating. This task emphasizes fine-grained semantic control and physical plausibility while capturing the rich diversity of daily interactions.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Trends and challenges in robot manipulation

URL: View paper

Brief Assessment

Robot Manipulation Trends[56] is a review paper discussing robotic manipulation challenges and progress, not a method for generating hand-object interactions. It focuses on robotic systems learning manipulation skills, not on synthesizing diverse human hand-object interactions including non-grasping actions.

2. O2o-afford: Annotation-free large-scale object-object affordance learning

URL: View paper

Brief Assessment

O2O-Afford[57] focuses on object-object interaction affordance learning (e.g., placing objects on tables, fitting objects in drawers), not hand-object interactions. The candidate addresses a fundamentally different problem domain involving interactions between two objects rather than hand-object manipulations like pushing, poking, or rotating with hands.

3. Push to know!-visuo-tactile based active object parameter inference with dual differentiable filtering

URL: View paper

Brief Assessment

Push to Know[58] focuses on estimating physical object parameters (mass, friction, center of mass, inertia) through non-prehensile pushing actions, not on generating diverse hand-object interactions or hand poses for various manipulation types.

4. Hand-object interaction: from grasping to using

URL: View paper

Brief Assessment

Grasping to Using[62] appears to focus on grasp synthesis algorithms and object manipulation tasks, but the provided context is too fragmentary to assess whether it addresses the same free-form interaction generation task with fine-grained semantic control proposed in the original paper.

5. Predictive visuo-tactile interactive perception framework for object properties inference

URL: View paper

Brief Assessment

Visuo-tactile Perception[55] focuses on robotic exploration of object properties (stiffness, mass, friction) through pushing and pulling actions, not on generating diverse hand-object interactions for synthesis purposes. The candidate addresses autonomous property inference for manipulation, while the original addresses controllable generation of interaction poses beyond grasping.

6. Push-grasping with dexterous hands: Mechanics and a method

URL: View paper

Brief Assessment

Push-grasping Dexterous[61] focuses on robotic manipulation mechanics for push-grasp planning with uncertainty handling, not on generating diverse hand-object interactions or non-grasping actions like the original paper's free-form interaction generation task.

7. Experimental Evaluation of Precise Placement with Pushing Primitive Based on Cartesian Force Control

URL: View paper

Brief Assessment

This candidate focuses on robotic manipulation using pushing primitives with force control for precise object placement in manufacturing. It does not address hand-object interaction generation, semantic control, or diverse non-grasping manipulations in the context of synthesis or generation tasks.

8. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions

URL: View paper

Brief Assessment

Diffh2o[9] focuses on generating hand-object interactions from textual descriptions but does not address the specific task of free-form HOI generation including non-grasping actions like pushing, poking, and rotating that the original paper emphasizes.

9. MultiSCOPE: Disambiguating in-hand object poses with proprioception and sequential interactions

URL: View paper

Brief Assessment

MultiSCOPE[60] focuses on state estimation for grasped object pose using proprioception and tactile feedback in robotic manipulation, not on generating diverse hand-object interactions from text descriptions. The candidate addresses pose estimation during manipulation, while the original paper proposes synthesizing varied interaction types (pushing, poking, rotating) from semantic descriptions.

10. Learning from human videos for robotic manipulation

URL: View paper

Brief Assessment

Learning Human Videos[59] focuses on learning robotic manipulation policies from human videos, not on generating diverse hand-object interactions including non-grasping actions. The candidate addresses policy learning for robot control, while the original paper proposes a generation task for synthesizing varied interaction poses.

Contribution 2: WildO2 dataset with automated construction pipeline

Description: The authors build WildO2, a large-scale 3D hand-object interaction dataset collected from in-the-wild videos. The dataset includes diverse non-grasping interactions with fine-grained semantic annotations, constructed through an automated pipeline that recovers 3D interactions from internet videos.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. H2o: Two hands manipulating objects for first person interaction recognition

URL: View paper

Brief Assessment

H2O[44] focuses on egocentric two-handed object manipulation with multi-view RGB-D capture in controlled settings, not on automated construction from in-the-wild internet videos with diverse non-grasping interactions as in WildO2.

2. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications

URL: View paper

Brief Assessment

Egocentric Hand Segmentation[41] focuses on 2D hand-object segmentation in egocentric videos with per-pixel labels, not 3D hand-object interaction reconstruction from in-the-wild videos. The candidate addresses a fundamentally different task (2D segmentation vs. 3D interaction recovery) and does not demonstrate prior work on automated pipelines for recovering 3D interactions from internet videos.

3. Understanding human hands in contact at internet scale

URL: View paper

Brief Assessment

Hands in Contact[43] focuses on 2D hand detection and contact state classification in internet videos, not 3D hand-object interaction reconstruction with semantic annotations. The candidate does not construct 3D interaction datasets from videos.

4. HO-Cap: A Capture System and Dataset for 3D Reconstruction and Pose Tracking of Hand-Object Interaction

URL: View paper

Brief Assessment

HO-Cap[47] focuses on a multi-camera capture system with semi-automatic annotation for controlled hand-object interaction videos, not on building large-scale datasets from in-the-wild internet videos with automated pipelines as in the original paper.

5. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions

URL: View paper

Brief Assessment

H+O[45] focuses on egocentric recognition of hand-object poses from RGB video for activity recognition tasks, not on building large-scale 3D interaction datasets from in-the-wild videos with automated pipelines and semantic annotations.

6. HOT3D: Hand and Object Tracking in 3D from Egocentric Multi-View Videos

URL: View paper

Brief Assessment

HOT3D[46] focuses on egocentric multi-view hand-object tracking with motion capture ground truth in controlled environments, while WildO2 addresses in-the-wild 3D hand-object interaction reconstruction from internet videos with semantic annotations. The datasets serve different purposes and use fundamentally different data collection methodologies.

7. HanDyVQA: A Video QA Benchmark for Fine-Grained Hand-Object Interaction Dynamics

URL: View paper

Brief Assessment

HanDyVQA[48] focuses on video question-answering for hand-object interaction understanding in egocentric videos, not on building 3D hand-object interaction datasets from in-the-wild videos with automated reconstruction pipelines. The candidate paper builds a benchmark for evaluating models' understanding of HOI dynamics through QA tasks, while the original paper constructs a 3D dataset with geometric annotations through automated 3D reconstruction from internet videos.

8. Hoi4d: A 4d egocentric dataset for category-level human-object interaction

URL: View paper

Brief Assessment

HOI4D[42] focuses on egocentric 4D video sequences with category-level object pose tracking and action segmentation, not on automated construction pipelines for recovering 3D interactions from in-the-wild internet videos. The datasets serve different purposes and use different data collection methodologies.

9. Affordance diffusion: Synthesizing hand-object interactions

URL: View paper

Brief Assessment

Affordance Diffusion[40] focuses on synthesizing hand-object interaction images from single RGB images using diffusion models, not on constructing large-scale 3D datasets from in-the-wild videos. Their dataset (HO3Pairs) is constructed differently - using inpainting to create paired data from egocentric videos rather than recovering 3D interactions with semantic annotations from internet videos.

10. RGBD Objects in the Wild: Scaling Real-World 3D Object Learning from RGB-D Videos

URL: View paper

Brief Assessment

RGBD Objects Wild[49] focuses on RGB-D object videos for 3D object learning tasks (novel view synthesis, pose estimation), not hand-object interaction datasets with semantic annotations for interaction intent modeling.

Contribution 3: TOUCH framework for controllable HOI generation

Description: The authors introduce TOUCH, a three-stage generation framework featuring explicit contact modeling, multi-level diffusion with coarse-to-fine semantic control, and physical constraint refinement. This framework enables the generation of diverse, controllable, and physically plausible hand-object interactions guided by fine-grained textual descriptions.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. A motion conditioned diffusion model for real-time hand trajectory semantic prediction

URL: View paper

Brief Assessment

Hand Trajectory Semantic[52] focuses on semantic prediction of hand trajectories using diffusion models for motion synthesis, not on controllable hand-object interaction generation with explicit contact modeling and multi-level semantic control as proposed in TOUCH.

2. Thor: Text to human-object interaction diffusion via relation intervention

URL: View paper

Brief Assessment

THOR[12] focuses on generating full-body human-object interactions with dynamic objects from text, while TOUCH addresses free-form hand-object interactions with explicit contact modeling and fine-grained semantic control. These are distinct problem scopes with different technical approaches.

3. Controllable human-object interaction synthesis

URL: View paper

Brief Assessment

Controllable HOI Synthesis[1] focuses on full-body human-object interaction synthesis in 3D scenes using waypoint-based path planning and language descriptions, while the original paper addresses hand-object interactions with fine-grained contact modeling and part-level hand control. The candidate does not challenge the novelty of the three-stage TOUCH framework with explicit contact modeling, multi-level diffusion, and physical constraint refinement for hand-centric interactions.

4. Hoidiffusion: Generating realistic 3d hand-object interaction data

URL: View paper

Brief Assessment

Hoidiffusion[3] focuses on generating 2D images of hand-object interactions using diffusion models conditioned on 3D geometric structures and text, primarily for data augmentation. The ORIGINAL paper addresses 3D hand pose generation with explicit contact modeling and multi-level semantic control for free-form interactions beyond grasping, representing a fundamentally different technical approach and scope.

5. Graspdiff: Grasping generation for hand-object interaction with multimodal guided diffusion

URL: View paper

Brief Assessment

GraspDiff[50] focuses specifically on grasping generation using diffusion models with contact affordance and image content as conditions, while the original paper addresses broader free-form hand-object interactions (including non-grasping actions like pushing, poking, rotating) with fine-grained textual control and multi-level semantic conditioning. The candidate's scope is limited to grasp-centric paradigms, not the diverse interaction types central to TOUCH.

6. Diffusion-guided reconstruction of everyday hand-object interaction clips

URL: View paper

Brief Assessment

Diffusion-guided Reconstruction[54] focuses on reconstructing hand-object interactions from video clips using diffusion models as priors for 3D geometry inference, not on generating controllable interactions from text descriptions. The tasks are fundamentally different: reconstruction from observed videos versus generation from semantic specifications.

7. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models

URL: View paper

Brief Assessment

Hoi-diff[6] focuses on generating full-body human-object interaction motions (sequences) with dynamic objects using text prompts, while TOUCH addresses static hand-object interactions with fine-grained contact control and physical constraints. The tasks and technical approaches differ fundamentally.

8. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions

URL: View paper

Brief Assessment

Diffh2o[9] proposes a temporal two-stage diffusion framework with grasp guidance, which differs from TOUCH's three-stage framework with explicit contact modeling, multi-level diffusion, and physical constraint refinement. The technical approaches are distinct.

9. Coda: Coordinated diffusion noise optimization for whole-body manipulation of articulated objects

URL: View paper

Brief Assessment

CODA[51] focuses on whole-body manipulation of articulated objects with coordinated body-hand motion, while the original paper addresses free-form hand-object interactions with fine-grained semantic control. CODA[51] does not address the specific novelty claims of explicit contact modeling for diverse non-grasping interactions or multi-level diffusion with coarse-to-fine semantic control for hand-only generation.

10. HunyuanVideo-HOMA: Generic Human-Object Interaction in Multimodal Driven Human Animation

URL: View paper

Brief Assessment

HunyuanVideo-HOMA[53] focuses on human animation driven by multimodal inputs (user-editable motion, object customization, sparse pose sequences, object trajectories, text prompts), not on hand-object interaction generation with diffusion models and semantic control as in TOUCH.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] TOUCH: Text-guided Controllable Generation of Free-Form Hand-Object Interactions View paper
- [1] Controllable human-object interaction synthesis View paper
- [2] TASTE-Rob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation View paper
- [3] Hoidiffusion: Generating realistic 3d hand-object interaction data View paper
- [4] SIGHT: Single-Image Conditioned Generation of Hand Trajectories for Hand-Object Interaction View paper
- [5] Chainhoi: Joint-based kinematic chain modeling for human-object interaction generation View paper
- [6] Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models View paper
- [7] Interdreamer: Zero-shot text to 3d dynamic human-object interaction View paper
- [8] Text-driven affordance learning from egocentric vision View paper
- [9] Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions View paper
- [10] Cg-hoi: Contact-guided 3d human-object interaction generation View paper
- [11] Human-object interaction from human-level instructions View paper
- [12] Thor: Text to human-object interaction diffusion via relation intervention View paper

- [13] OOD-HOI: Text-driven 3d whole-body human-object interactions generation beyond training domains View paper
- [14] InteractAnything: Zero-shot Human Object Interaction Synthesis via LLM Feedback and Object Affordance Parsing View paper
- [15] Hoianimator: Generating text-prompt human-object animations using novel perceptive diffusion models View paper
- [16] HOIGPT: Learning Long-Sequence Hand-Object Interaction with Language Models View paper
- [17] Efficient Explicit Joint-level Interaction Modeling with Mamba for Text-guided HOI Generation View paper
- [18] HOIDiNi: Human-Object Interaction through Diffusion Noise Optimization View paper
- [19] Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions View paper
- [20] Generating human interaction motions in scenes with text control View paper
- [21] Multimodal human-intent modeling for contextual robot-to-human handovers of arbitrary objects View paper
- [22] Ram: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation View paper
- [23] OpenHOI: Open-World Hand-Object Interaction Synthesis with Multimodal Large Language Model View paper
- [24] Text2HOI: Text-Guided 3D Motion Generation for Hand-Object Interaction View paper
- [25] Prompting future driven diffusion model for hand motion prediction View paper
- [26] MEgoHand: Multimodal Egocentric Hand-Object Interaction Motion Generation View paper
- [27] TIGeR: Text-Instructed Generation and Refinement for Template-Free Hand-Object Interaction View paper
- [28] Nl2contact: Natural language guided 3d hand-object contact modeling with diffusion model View paper
- [29] Himo: A new benchmark for full-body human interacting with multiple objects View paper
- [30] Textim: Part-aware interactive motion synthesis from text View paper
- [31] OASIS: Object-guided Attention for Text-conditional Diffusion Synthesis of Human Interaction Sequences View paper
- [32] InterFusion: Text-Driven Generation of 3D Human-Object Interaction View paper
- [33] InteractMove: Text-Controlled Human-Object Interaction Generation in 3D Scenes with Movable Objects View paper
- [34] SIGHT: Synthesizing Image-Text Conditioned and Geometry-Guided 3D Hand-Object Trajectories View paper
- [35] EigenActor: Variant Body-Object Interaction Generation Evolved from Invariant Action Basis Reasoning View paper
- [36] FunHOI: Annotation-Free 3D Hand-Object Interaction Generation via Functional Text Guidanc View paper
- [37] DHAGrasp: Synthesizing Affordance-Aware Dual-Hand Grasps with Text Instructions View paper
- [38] Towards Semantic 3D Hand-Object Interaction Generation via Functional Text Guidance View paper
- [39] Prompt-Propose-Verify: A Reliable Hand-Object-Interaction Data Generation Framework using Foundational Models View paper
- [40] Affordance diffusion: Synthesizing hand-object interactions View paper
- [41] Fine-grained egocentric hand-object segmentation: Dataset, model, and applications View paper
- [42] Hoi4d: A 4d egocentric dataset for category-level human-object interaction View paper
- [43] Understanding human hands in contact at internet scale View paper
- [44] H2o: Two hands manipulating objects for first person interaction recognition View paper
- [45] H+ o: Unified egocentric recognition of 3d hand-object poses and interactions View paper
- [46] HOT3D: Hand and Object Tracking in 3D from Egocentric Multi-View Videos View paper
- [47] HO-Cap: A Capture System and Dataset for 3D Reconstruction and Pose Tracking of Hand-Object Interaction View paper
- [48] HanDyVQA: A Video QA Benchmark for Fine-Grained Hand-Object Interaction Dynamics View paper
- [49] RGBD Objects in the Wild: Scaling Real-World 3D Object Learning from RGB-D Videos View paper
- [50] Graspdiff: Grasping generation for hand-object interaction with multimodal guided diffusion View paper
- [51] Coda: Coordinated diffusion noise optimization for whole-body manipulation of articulated objects View paper
- [52] A motion conditioned diffusion model for real-time hand trajectory semantic prediction View paper
- [53] HunyuanVideo-HOMA: Generic Human-Object Interaction in Multimodal Driven Human Animation View paper
- [54] Diffusion-guided reconstruction of everyday hand-object interaction clips View paper
- [55] Predictive visuo-tactile interactive perception framework for object properties inference View paper
- [56] Trends and challenges in robot manipulation View paper
- [57] O2o-afford: Annotation-free large-scale object-object affordance learning View paper
- [58] Push to know!-visuo-tactile based active object parameter inference with dual differentiable filtering View paper
- [59] Learning from human videos for robotic manipulation View paper
- [60] MultiSCOPE: Disambiguating in-hand object poses with proprioception and sequential interactions View paper
- [61] Push-grasping with dexterous hands: Mechanics and a method View paper
- [62] Hand-object interaction: from grasping to using View paper
- [63] Experimental Evaluation of Precise Placement with Pushing Primitive Based on Cartesian Force Control View paper