Novelty Assessment Report

Paper: Human-Object Interaction via Automatically Designed VLM-Guided Motion Policy

PDF URL: https://openreview.net/pdf?id=LfkPlFTfe0

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-27

Abstract

Human-object interaction (HOI) synthesis is crucial for applications in animation, simulation, and robotics. However, existing approaches either rely on expensive motion capture data or require manual reward engineering, limiting their scalability and generalizability. In this work, we introduce the first unified physics-based HOI framework that leverages Vision-Language Models (VLMs) to enable long-horizon interactions with diverse object types — including static, dynamic, and articulated objects. We introduce VLM-Guided Relative Movement Dynamics (RMD), a fine-grained spatio-temporal bipartite representation that automatically constructs goal states and reward functions for reinforcement learning. By encoding structured relationships between human and object parts, RMD enables VLMs to generate semantically grounded, interaction-aware motion guidance without manual reward tuning. To support our methodology, we present Interplay, a novel dataset with thousands of long-horizon static and dynamic interaction plans. Extensive experiments demonstrate that our framework outperforms existing methods in synthesizing natural, human-like motions across both simple single-task and complex multi-task scenarios. For more details, please refer to our project webpage: https://vlm-rmd.github.io/.

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: **Synthesizing Physics-Based Human-Object Interactions**

A total of **50 papers** were analyzed and organized into a taxonomy with **32 categories**.

Taxonomy Overview

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

- **Physics-Based Motion Imitation and Control**
- **Kinematic and Diffusion-Based Synthesis**
- **Data-Driven Interaction Modeling and Datasets**
- **Scene-Aware Interaction Synthesis**
- **Language and Vision-Guided Interaction**
- **Physical Consistency and Contact Refinement**
- **Specialized Interaction Domains**
- **Physics Learning and Reasoning Foundations**

Complete Taxonomy Tree

- Synthesizing Physics-Based Human-Object Interactions Survey Taxonomy
- Physics-Based Motion Imitation and Control
  - Whole-Body Dynamic Interaction Imitation (2 papers)
  - [3] Intermimic: Towards universal whole-body control for physics-based human-object interactions (Xu, 2025) View paper
  - [5] Physhoi: Physics-based imitation of dynamic human-object interaction (Wang Yinhuai, 2023) View paper
  - Hand-Object Manipulation Control (2 papers)
  - [8] D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions (Christen, 2022) View paper
  - [31] Hand-object interaction controller (hoic): Deep reinforcement learning for reconstructing interactions with physics (Haoyu Hu, 2024) View paper
  - Multi-Skill Unified Control Policies (1 papers)
  - [14] TokenHSI: Unified Synthesis of Physical Human-Scene Interactions through Task Tokenization (Pan Liang, 2025) View paper
  - Hierarchical Full-Body Hand-Object Synthesis (1 papers)
  - [17] Physically plausible full-body hand-object interaction synthesis (Jona Braun, 2024) View paper
- Kinematic and Diffusion-Based Synthesis
  - Physics-Aware Kinematic Synthesis (2 papers)
  - [1] Force: Physics-aware human-object interaction (Zhang Xiaohan, 2025) View paper
  - [35] Force: Dataset and method for intuitive physics guided human-object interaction (Xiaohan Zhang, 2024) View paper
  - Text-Driven Diffusion Synthesis (3 papers)
  - [4] Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models (Xiaogang Peng, 2025) View paper
  - [6] Interdiff: Generating 3d human-object interactions with physics-informed diffusion (Sirui Xu, 2023) View paper
  - [18] OOD-HOI: Text-driven 3d whole-body human-object interactions generation beyond training domains (Zhang Yi-xuan, 2024) View paper
  - Noise Optimization for Interaction Synthesis (2 papers)
  - [21] Coda: Coordinated diffusion noise optimization for whole-body manipulation of articulated objects (Pi, 2025) View paper
  - [44] HOIDiNi: Human-Object Interaction through Diffusion Noise Optimization (Tevet, 2025) View paper
  - Hand Manipulation Trajectory Generation (2 papers)

  - [30] SIGHT: Synthesizing Image-Text Conditioned and Geometry-Guided 3D Hand-Object Trajectories (Gavryushin, 2025) View paper
  - [38] ManiDext: Hand-Object Manipulation Synthesis via Continuous Correspondence Embeddings and Residual-Guided Diffusion (Jiajun Zhang, 2024) View paper
- Data-Driven Interaction Modeling and Datasets
  - Full-Body Articulated Interaction Datasets (2 papers)
  - [2] Full-body articulated human-object interaction (Nan Jiang, 2023) View paper
  - [29] Interact: Advancing large-scale versatile 3d human-object interaction generation (Xu, 2025) View paper
  - Interaction-Preserving Motion Retargeting (1 papers)
  - [16] Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction (Yang, 2025) View paper
  - Synthetic Data Generation for Interactions (2 papers)
  - [13] Physics-driven data generation for contact-rich manipulation via trajectory optimization (Yang, 2025) View paper
  - [33] VR-based generation of photorealistic synthetic data for training hand-object tracking models (Chengyan Zhang, 2024) View paper
  - Collaborative and Multi-Human Interaction Datasets (2 papers)
  - [11] Coohoi: Learning cooperative human-object interaction with manipulated object dynamics (Jinkun Cao, 2024) View paper
  - [24] Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement (Yun Liu, 2025) View paper
- Scene-Aware Interaction Synthesis
  - Human Placement in Static Scenes (2 papers)
  - [9] PhySIC: Physically Plausible 3D Human-Scene Interaction and Contact from a Single Image (Xue, 2025) View paper
  - [27] Generating 3D People in Scenes Without People (Yan Zhang, 2019) View paper
  - Dynamic Scene Interaction Synthesis (1 papers)
  - [25] Synthesizing Physical Character-Scene Interactions (Mohamed Hassan, 2023) View paper
  - Pick-and-Place in Cluttered Environments (1 papers)
  - [26] CHOICE: Coordinated human-object interaction in cluttered environments for pick-and-place actions (Jintao Lu, 2024) View paper
  - Scene Layout Generation from Motion (1 papers)
  - [45] Physics-based scene layout generation from human motion (Li Jianan, 2024) View paper
  - Physically Interactable Scene Synthesis (1 papers)
  - [19] Physcene: Physically interactable 3d scene synthesis for embodied ai (Yandan Yang, 2024) View paper
- Language and Vision-Guided Interaction
  - VLM-Guided Motion Policy Design ★ (2 papers)
  - [0] Human-Object Interaction via Automatically Designed VLM-Guided Motion Policy (Anon et al., 2026) View paper
  - [36] Human-Object Interaction with Vision-Language Model Guided Relative Movement Dynamics (Zekai Deng, 2025) View paper
  - LLM-Driven Task Planning for Manipulation (1 papers)
  - [10] Human-object interaction from human-level instructions (Wu Zhen, 2025) View paper
  - Text-to-3D Interaction Generation (1 papers)
  - [41] Interdreamer: Zero-shot text to 3d dynamic human-object interaction (Liang-Yan Gui, 2024) View paper
  - Language-Guided Sparse Control (1 papers)
  - [15] Controllable human-object interaction synthesis (Jiaman Li, 2024) View paper
  - Contact-Aware Text-Driven Motion (1 papers)
  - [40] Contact-aware Human Motion Generation from Textual Descriptions (Ma, 2024) View paper
- Physical Consistency and Contact Refinement
  - Differentiable Physics for Contact Reasoning (1 papers)
  - [32] HandyPriors: Physically Consistent Perception of Hand-Object Interactions with Differentiable Priors (Zhang Shutong, 2023) View paper
  - Interaction Denoising and Artifact Removal (1 papers)
  - [12] Physics-aware hand-object interaction denoising (Haowen Luo, 2024) View paper
  - Interactive Motion Editing with Physics (1 papers)
  - [28] Interactive synthesis of human-object interaction (Sumit Jain, 2009) View paper
- Specialized Interaction Domains
  - Video Generation with Physical Dynamics (2 papers)
  - [7] PhysDreamer: Physics-Based Interaction with 3D Objects via Video Generation (Tianyuan ZHANG, 2024) View paper
  - [23] PhysMaster: Mastering Physical Representation for Video Generation via Reinforcement Learning (Chen Xi, 2025) View paper
  - Robotic Manipulation Planning (4 papers)
  - [20] Differentiable physics and stable modes for tool-use and manipulation planning (Marc Toussaint, 2018) View paper
  - [43] Describing Physics For Physical Reasoning: Force-Based Sequential Manipulation Planning (Marc Toussaint, 2020) View paper
  - [47] RAIL: Robot Affordance Imagination with Large Language Models (Zhang Ceng, 2024) View paper
  - [50] Learning Robust Grasping Strategy Through Tactile Sensing and Adaption Skill (Yueming Hu, 2024) View paper
  - Immersive VR/AR Interaction Systems (2 papers)
  - [34] Towards immersive human-x interaction: A real-time framework for physically plausible motion synthesis (Shi Ye, 2025) View paper
  - [39] A Mixed Reality Training System for Hand-Object Interaction in Simulated Microgravity Environments (Kanglei Zhou, 2023) View paper
  - Photorealistic Rendering of Interactions (3 papers)
  - [22] Feature Splatting: Language-Driven Physics-Based Scene Synthesis and Editing (Ri-Zhao Qiu, 2024) View paper
  - [37] GraspDiffusion: Synthesizing Realistic Whole-body Hand-Object Interaction (Kwon, 2024) View paper
  - [42] GASPACHO: Gaussian Splatting for Controllable Humans and Objects (Mir, 2025) View paper
  - Kinematic Models with Physical Interactions (1 papers)
  - [46] Half-Physics: Enabling Kinematic 3D Human Model with Physical Interactions (Siyao Li, 2025) View paper
- Physics Learning and Reasoning Foundations
  - Neural Physics Simulation (1 papers)

- ◦ [49] Interaction networks for learning about objects, relations and physics (Battaglia, 2016) View paper
- ◦ Intuitive Physics for Decision Making (1 papers)
- ◦ [48] Learning to Play Video Games with Intuitive Physics Priors (Jaiswal Abhishek, 2024) View paper

Narrative

Core task: synthesizing physics-based human-object interactions. The field organizes around several complementary branches that address different facets of generating realistic human motions with objects. Physics-Based Motion Imitation and Control emphasizes reinforcement learning and trajectory optimization to produce dynamically stable behaviors, often drawing on reference motion data or learned policies (e.g., Intermimic Universal Control[3], PhysHOI Imitation[5]). Kinematic and Diffusion-Based Synthesis leverages generative models—particularly diffusion frameworks—to sample plausible interaction sequences while balancing kinematic realism with physical constraints (HOI Diff[4], InterDiff Physics[6]). Data-Driven Interaction Modeling and Datasets focuses on curating large-scale collections and benchmarks that capture diverse contact patterns and object affordances. Scene-Aware Interaction Synthesis tackles the challenge of placing and adapting motions within cluttered or geometrically complex environments, while Language and Vision-Guided Interaction explores how high-level instructions or visual cues can steer motion policies. Physical Consistency and Contact Refinement refines generated outputs to satisfy contact mechanics and force balance, and Specialized Interaction Domains targets niche settings such as hand manipulation or cooperative tasks.

Recent work highlights a tension between purely kinematic generation—which can produce visually smooth results quickly—and physics-driven approaches that enforce dynamic feasibility at the cost of greater computational expense. A growing number of studies blend diffusion priors with physics-based post-processing to achieve both diversity and stability (Physics Aware Denoising[12], Physics Driven Generation[13]). Within the Language and Vision-Guided Interaction branch, VLM Guided Motion[0] exemplifies efforts to integrate vision-language models into motion policy design, enabling more intuitive control through natural language or image-based prompts. This direction contrasts with purely data-driven methods like Full Body HOI[2] or force-centric frameworks such as Force Physics HOI[1], which prioritize contact realism over high-level semantic guidance. By situating language-conditioned policies alongside physics simulators, VLM Guided Motion[0] bridges the gap between user intent and physically grounded execution, a theme echoed by neighboring work on relative movement reasoning (VLM Relative Movement[36]).

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Human-Object Interaction with Vision-Language Model Guided Relative Movement Dynamics

Authors: Zekai Deng, Ye Shi, Kaiyang Ji, Lan Xu, Shaoli Huang, et al. (6 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

N/A

⚠ Similarity Notice

The titles are nearly identical, both focusing on VLM-guided motion policy design for human-object interactions using relative movement dynamics. The candidate paper's title appears to be a slight variation of the original paper's title, and both describe the same core technical approach of using vision-language models to automatically design reward functions and goal states through a Relative Movement Dynamics (RMD) representation. This strongly suggests they are the same paper or very close variants.

Contributions Analysis

Overall novelty summary. The paper proposes a unified physics-based framework that uses Vision-Language Models to guide reinforcement learning for long-horizon human-object interactions across static, dynamic, and articulated objects. It resides in the 'VLM-Guided Motion Policy Design' leaf under 'Language and Vision-Guided Interaction', which contains only two papers total (including this one). This represents a relatively sparse research direction within the broader taxonomy of 50 papers across 32 leaf nodes, suggesting the specific combination of VLMs with physics-based HOI synthesis is an emerging area rather than a crowded subfield.

The parent branch 'Language and Vision-Guided Interaction' encompasses five leaves addressing different aspects of language-conditioned synthesis: VLM-guided policy design, LLM-driven task planning, text-to-3D generation, language-guided sparse control, and contact-aware text-driven motion. Neighboring branches include 'Physics-Based Motion Imitation and Control' (which emphasizes learning from motion capture without language guidance) and 'Kinematic and Diffusion-Based Synthesis' (which uses generative models rather than reinforcement learning). The taxonomy's scope notes clarify that VLM-guided methods specifically automate reward design, distinguishing them from manual reward engineering approaches in adjacent physics-based leaves.

Among 29 candidates examined through semantic search and citation expansion, none were found to clearly refute any of the three main contributions. The first contribution (unified VLM-physics framework) examined 10 candidates with zero refutations; the second (RMD representation) examined 9 with zero refutations; the third (Interplay dataset) examined 10 with zero refutations. This limited search scope—covering roughly 60% of the taxonomy's total papers—suggests that within the examined literature, the specific integration of VLMs for automatic reward construction in physics-based HOI appears relatively unexplored, though the analysis cannot claim exhaustive coverage of all potentially relevant prior work.

The analysis indicates novelty within the examined scope, particularly in combining VLM-based semantic reasoning with physics simulation for diverse object types. However, the search examined only top-K semantic matches rather than a comprehensive field survey, and the sparse population of the target taxonomy leaf (2 papers) may reflect either genuine novelty or incomplete taxonomy coverage. The contribution-level statistics consistently show no clear refutations, but this should be interpreted as 'no overlapping work found among 29 candidates' rather than definitive proof of absolute novelty across the entire research landscape.

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

Contribution 1: Unified physics-based HOI framework leveraging VLMs for long-horizon interactions

Description: The authors introduce a unified framework that uses Vision-Language Models to enable physics-based synthesis of long-horizon human-object interactions. This framework supports diverse object types (static, dynamic, and articulated) without requiring expensive motion capture data or manual reward engineering.

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. Grounding 3D Object Affordance with Language Instructions, Visual Observations and Interactions

URL: View paper

Brief Assessment

Grounding Affordance Language[66] focuses on 3D object affordance grounding for manipulation tasks using vision-language models, not on physics-based synthesis of long-horizon human-object interactions or motion policy learning through reinforcement learning.

2. Human-object interaction from human-level instructions

URL: View paper

Brief Assessment

Human Level Instructions[10] focuses on object manipulation tasks with detailed finger movements using LLMs for planning and diffusion models for motion generation, rather than a unified physics-based RL framework with VLM-guided reward design for diverse object types as proposed in the original paper.

3. Controllable human-object interaction synthesis

URL: View paper

Brief Assessment

Controllable HOI Synthesis[15] focuses on kinematic motion synthesis using diffusion models for human-object interactions, not physics-based reinforcement learning frameworks. The candidate does not address physics simulation, reward engineering, or RL-based policy learning that are central to the original contribution.

4. HumanVLA: Towards Vision-Language Directed Object Rearrangement by Physical Humanoid

URL: View paper

Brief Assessment

HumanVLA[67] focuses on vision-language directed object rearrangement using a teacher-student distillation framework, not on VLM-guided motion policy design with automatic reward construction for diverse interaction types.

5. OpenHOI: Open-World Hand-Object Interaction Synthesis with Multimodal Large Language Model

URL: View paper

Brief Assessment

OpenHOI[60] focuses on hand-object interactions with affordance grounding and diffusion-based synthesis, while the original paper addresses full-body human-object interactions using physics-based RL with VLM-guided reward design. These are fundamentally different problem domains and technical approaches.

6. Generating Human Motion in 3D Scenes from Text Descriptions

URL: View paper

Brief Assessment

Motion Text Scenes[63] focuses on generating human motions in 3D scenes from text descriptions using diffusion models and object-centric representations, not on physics-based synthesis with VLM-guided reinforcement learning for diverse object interactions.

7. Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation

URL: View paper

Brief Assessment

Sofar Orientation[62] focuses on semantic orientation for spatial reasoning and 6-dof manipulation tasks, not on physics-based synthesis of long-horizon human-object interactions using VLMs.

8. PhyGrasp: Generalizing Robotic Grasping with Physics-informed Large Multimodal Models

URL: View paper

Brief Assessment

PhyGrasp[65] focuses on robotic grasping using multimodal models for object manipulation, not human-object interaction synthesis or long-horizon motion generation for humanoid characters.

9. AffordanceLLM: Grounding Affordance from Vision Language Models

URL: View paper

Brief Assessment

AffordanceLLM[61] focuses on affordance grounding (identifying interaction regions in images) rather than physics-based motion synthesis. It does not address long-horizon interaction synthesis, reinforcement learning for motion policies, or physics simulation of human-object interactions.

10. Anyskill: Learning open-vocabulary physical skill for interactive agents

URL: View paper

Brief Assessment

AnySkill[64] focuses on learning open-vocabulary physical skills through hierarchical control (low-level controller + high-level policy) using CLIP-based image rewards, not on VLM-guided planning for long-horizon human-object interactions with diverse object types as in the original paper.

Contribution 2: VLM-Guided Relative Movement Dynamics (RMD) representation

Description: The authors propose RMD, a structured spatio-temporal representation that encodes fine-grained relationships between human and object parts. This representation enables VLMs to automatically generate goal states and reward functions for reinforcement learning, eliminating the need for manual reward engineering while supporting both static and dynamic interactions.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. GENFLOWRL: Generative Object-Centric Flow Matching for Reward Shaping in Visual Reinforcement Learning

URL: View paper

Brief Assessment

GENFLOWRL[53] focuses on object-centric flow representations for reward shaping in visual RL, not on VLM-guided spatial-temporal representations for automatic reward generation in human-object interaction. The candidate addresses robot manipulation tasks using generated flow matching, while the original paper proposes RMD for physics-based human-object interaction synthesis with VLM guidance.

2. Learning Reward Functions for Robotic Manipulation by Observing Humans

URL: View paper

Brief Assessment

Learning Reward Observation[59] focuses on learning reward functions from human demonstration videos for robotic manipulation, using temporal regression and time-contrastive learning in embedding spaces. This differs fundamentally from the original paper's RMD, which is a structured bipartite graph representation encoding fine-grained spatial-temporal relationships between human and object parts for VLM-guided motion synthesis.

3. Human-oriented representation learning for robotic manipulation

URL: View paper

Brief Assessment

Human Oriented Representation[51] focuses on learning visual representations for robotic manipulation through multi-task perceptual skills (hand detection, state estimation) rather than automatic reward function generation. The candidate does not address VLM-guided goal state construction or reward function design for reinforcement learning.

4. Teaching Virtual Agents to Perform Complex Spatial-Temporal Activities.

URL: View paper

Brief Assessment

Virtual Agents Activities[58] focuses on teaching virtual agents spatial-temporal activities through qualitative spatial reasoning (QSR) and reinforcement learning from motion capture data. While both papers address spatial-temporal representations, the candidate uses QSR calculi (cardinal direction, qualitative distance/trajectory) for discrete action recognition in 2D block-world tasks, whereas the original proposes RMD as a fine-grained bipartite graph encoding part-level human-object dynamics with VLM-guided automatic reward generation for physics-based 3D interaction synthesis.

5. Deep selective feature learning for action recognition

URL: View paper

Brief Assessment

Selective Feature Learning[54] focuses on deep learning for action recognition in videos, not on spatial-temporal representations for automatic reward function generation in human-object interaction or reinforcement learning frameworks.

6. Task-Oriented Scanpath Prediction with Spatial-Temporal Information in Driving Scenarios

URL: View paper

Brief Assessment

Scanpath Driving[52] focuses on predicting human visual attention patterns (scanpaths) in driving scenarios, not on designing reward functions for human-object interaction synthesis. The spatial-temporal information in the candidate relates to eye movement trajectories, not to structured relationships between human and object parts for RL-based motion generation.

7. LSTM-GCN Hybrid Architecture for Model Predictive Control of Deformable Linear Objects

URL: View paper

Brief Assessment

LSTM-GCN Deformable[56] focuses on model predictive control of deformable linear objects using LSTM-GCN hybrid architectures for spatiotemporal feature fusion. This is fundamentally different from the original paper's VLM-guided RMD representation for automatic reward function generation in human-object interaction tasks.

8. A Survey on Reinforcement Learning of Vision-Language-Action Models for Robotic Manipulation

URL: View paper

Brief Assessment

VLA Survey[55] is a survey paper on vision-language-action models for robotic manipulation, not a research contribution proposing spatial-temporal representations for reward generation. The retrieved context fragments are insufficient to establish any substantive technical overlap with RMD.

9. Learning human utility from video demonstrations for deductive planning in robotics

URL: View paper

Brief Assessment

Learning Utility Video[57] focuses on learning utility functions from human preferences to guide deductive planning, using fluents to represent state changes. It does not propose a structured spatio-temporal representation for automatic reward function generation in VLM-guided RL frameworks.

Contribution 3: Interplay dataset for long-horizon static and dynamic interaction tasks

Description: The authors present Interplay, a new dataset containing thousands of long-horizon interaction plans that include both static and dynamic interaction tasks across varied scene contexts. This dataset addresses the gap in existing datasets that typically focus on either static interactions or object rearrangement separately.

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. Full-body articulated human-object interaction

URL: View paper

Brief Assessment

Full Body HOI[2] focuses on full-body articulated human-object interactions with sittable objects (chairs, sofas) captured via motion capture, not on long-horizon task planning with both static and dynamic objects across varied scene contexts as described in the original paper's Interplay dataset.

2. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation

URL: View paper

Brief Assessment

FurnitureBench[71] focuses on furniture assembly tasks with robotic manipulation, not human-object interaction datasets. The candidate addresses robotic assembly benchmarks rather than human motion interaction planning datasets.

3. Physhoi: Physics-based imitation of dynamic human-object interaction

URL: View paper

Brief Assessment

PhysHOI Imitation[5] introduces the BallPlay dataset focused specifically on basketball skills with high-dynamic contact scenarios, not general long-horizon static and dynamic interaction tasks across varied scene contexts as in the original paper's Interplay dataset.

4. Interdreamer: Zero-shot text to 3d dynamic human-object interaction

URL: View paper

Brief Assessment

InterDreamer[41] focuses on zero-shot text-to-3D human-object interaction generation without paired text-interaction training data, not on dataset construction for long-horizon tasks. The paper mentions existing datasets (BEHAVE, OMOMO, CHAIRS) but does not present a new dataset comparable to Interplay.

5. Dynamem: Online dynamic spatio-semantic memory for open world mobile manipulation

URL: View paper

Brief Assessment

DynaMem[74] focuses on dynamic spatio-semantic memory for mobile manipulation in changing environments, not on datasets for human-object interaction tasks. The candidate addresses environment mapping and object localization for robotic navigation, which is a fundamentally different problem domain from the original paper's focus on synthesizing human motion for interaction with objects.

6. Spatial-temporal human-object interaction detection

URL: View paper

Brief Assessment

Spatial Temporal Detection[68] focuses on video-based human-object interaction detection with spatial-temporal trajectories, not on long-horizon interaction planning datasets. The candidate constructs VIDOR-HOID for detection evaluation, while the original presents Interplay for planning tasks.

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

URL: View paper

Brief Assessment

HOI4D[69] focuses on egocentric category-level human-object interaction with 4D visual data (RGB-D sequences), not on long-horizon task planning with static and dynamic objects as emphasized in the original paper's Interplay dataset.

8. Exploring spatioâ□□temporal graph convolution for video-based humanâ□□object interaction recognition

URL: View paper

Brief Assessment

Spatio Temporal Graph[72] focuses on video-based human-object interaction recognition using graph convolutions on existing datasets (CAD-120, Something-Else, Charades), not on creating datasets for long-horizon interaction planning tasks.

9. Discovering a variety of objects in spatio-temporal human-object interactions

URL: View paper

Brief Assessment

Discovering Objects Interactions[73] focuses on third-view whole body-object interaction detection with diverse object discovery (1,000+ classes), not on long-horizon interaction planning with both static and dynamic tasks as in Interplay.

10. Interacted object grounding in spatio-temporal human-object interactions

URL: View paper

Brief Assessment

Interacted Object Grounding[70] focuses on grounding diverse interacted objects in third-person videos with 1,098 object classes, not on long-horizon interaction planning with both static and dynamic tasks as in the original paper's Interplay dataset.

Appendix: Text Similarity Detection

Textual similarity detection checked 30 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Human-Object Interaction with Vision-Language Model Guided Relative Movement Dynamics

Detected in: Core Task (sibling)

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Human-Object Interaction via Automatically Designed VLM-Guided Motion Policy View paper
- [1] Force: Physics-aware human-object interaction View paper
- [2] Full-body articulated human-object interaction View paper
- [3] Intermimic: Towards universal whole-body control for physics-based human-object interactions View paper
- [4] Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models View paper
- [5] Physhoi: Physics-based imitation of dynamic human-object interaction View paper
- [6] Interdiff: Generating 3d human-object interactions with physics-informed diffusion View paper
- [7] PhysDreamer: Physics-Based Interaction with 3D Objects via Video Generation View paper
- [8] D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions View paper
- [9] PhySIC: Physically Plausible 3D Human-Scene Interaction and Contact from a Single Image View paper
- [10] Human-object interaction from human-level instructions View paper
- [11] Coohoi: Learning cooperative human-object interaction with manipulated object dynamics View paper

- [12] Physics-aware hand-object interaction denoising View paper
- [13] Physics-driven data generation for contact-rich manipulation via trajectory optimization View paper
- [14] TokenHSI: Unified Synthesis of Physical Human-Scene Interactions through Task Tokenization View paper
- [15] Controllable human-object interaction synthesis View paper
- [16] Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction View paper
- [17] Physically plausible full-body hand-object interaction synthesis View paper
- [18] OOD-HOI: Text-driven 3d whole-body human-object interactions generation beyond training domains View paper
- [19] Physcene: Physically interactable 3d scene synthesis for embodied ai View paper
- [20] Differentiable physics and stable modes for tool-use and manipulation planning View paper
- [21] Coda: Coordinated diffusion noise optimization for whole-body manipulation of articulated objects View paper
- [22] Feature Splatting: Language-Driven Physics-Based Scene Synthesis and Editing View paper
- [23] PhysMaster: Mastering Physical Representation for Video Generation via Reinforcement Learning View paper
- [24] Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement View paper
- [25] Synthesizing Physical Character-Scene Interactions View paper
- [26] CHOICE: Coordinated human-object interaction in cluttered environments for pick-and-place actions View paper
- [27] Generating 3D People in Scenes Without People View paper
- [28] Interactive synthesis of human-object interaction View paper
- [29] Interact: Advancing large-scale versatile 3d human-object interaction generation View paper
- [30] SIGHT: Synthesizing Image-Text Conditioned and Geometry-Guided 3D Hand-Object Trajectories View paper
- [31] Hand-object interaction controller (hoic): Deep reinforcement learning for reconstructing interactions with physics View paper
- [32] HandyPriors: Physically Consistent Perception of Hand-Object Interactions with Differentiable Priors View paper
- [33] VR-based generation of photorealistic synthetic data for training hand-object tracking models View paper
- [34] Towards immersive human-x interaction: A real-time framework for physically plausible motion synthesis View paper
- [35] Force: Dataset and method for intuitive physics guided human-object interaction View paper
- [36] Human-Object Interaction with Vision-Language Model Guided Relative Movement Dynamics View paper
- [37] GraspDiffusion: Synthesizing Realistic Whole-body Hand-Object Interaction View paper
- [38] ManiDext: Hand-Object Manipulation Synthesis via Continuous Correspondence Embeddings and Residual-Guided Diffusion View paper
- [39] A Mixed Reality Training System for Hand-Object Interaction in Simulated Microgravity Environments View paper
- [40] Contact-aware Human Motion Generation from Textual Descriptions View paper
- [41] Interdreamer: Zero-shot text to 3d dynamic human-object interaction View paper
- [42] GASPACHO: Gaussian Splatting for Controllable Humans and Objects View paper
- [43] Describing Physics For Physical Reasoning: Force-Based Sequential Manipulation Planning View paper
- [44] HOIDiNi: Human-Object Interaction through Diffusion Noise Optimization View paper
- [45] Physics-based scene layout generation from human motion View paper
- [46] Half-Physics: Enabling Kinematic 3D Human Model with Physical Interactions View paper
- [47] RAIL: Robot Affordance Imagination with Large Language Models View paper
- [48] Learning to Play Video Games with Intuitive Physics Priors View paper
- [49] Interaction networks for learning about objects, relations and physics View paper
- [50] Learning Robust Grasping Strategy Through Tactile Sensing and Adaption Skill View paper
- [51] Human-oriented representation learning for robotic manipulation View paper
- [52] Task-Oriented Scanpath Prediction with Spatial-Temporal Information in Driving Scenarios View paper
- [53] GENFLOWRL: Generative Object-Centric Flow Matching for Reward Shaping in Visual Reinforcement Learning View paper
- [54] Deep selective feature learning for action recognition View paper
- [55] A Survey on Reinforcement Learning of Vision-Language-Action Models for Robotic Manipulation View paper
- [56] LSTM-GCN Hybrid Architecture for Model Predictive Control of Deformable Linear Objects View paper
- [57] Learning human utility from video demonstrations for deductive planning in robotics View paper
- [58] Teaching Virtual Agents to Perform Complex Spatial-Temporal Activities. View paper
- [59] Learning Reward Functions for Robotic Manipulation by Observing Humans View paper
- [60] OpenHOI: Open-World Hand-Object Interaction Synthesis with Multimodal Large Language Model View paper
- [61] AffordanceLLM: Grounding Affordance from Vision Language Models View paper
- [62] Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation View paper
- [63] Generating Human Motion in 3D Scenes from Text Descriptions View paper
- [64] Anyskill: Learning open-vocabulary physical skill for interactive agents View paper
- [65] PhyGrasp: Generalizing Robotic Grasping with Physics-informed Large Multimodal Models View paper
- [66] Grounding 3D Object Affordance with Language Instructions, Visual Observations and Interactions View paper
- [67] HumanVLA: Towards Vision-Language Directed Object Rearrangement by Physical Humanoid View paper
- [68] Spatial-temporal human-object interaction detection View paper
- [69] Hoi4d: A 4d egocentric dataset for category-level human-object interaction View paper
- [70] Interacted object grounding in spatio-temporal human-object interactions View paper
- [71] Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation View paper
- [72] Exploring spatioâ□□temporal graph convolution for video-based humanâ□□object interaction recognition View paper
- [73] Discovering a variety of objects in spatio-temporal human-object interactions View paper
- [74] Dynamem: Online dynamic spatio-semantic memory for open world mobile manipulation View paper