Novelty Assessment Report

Paper: Target-Aware Video Diffusion Models
PDF URL: https://openreview.net/pdf?id=311AxWM8FU
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

Abstract

We present a target-aware video diffusion model that generates videos from an input image, in which an actor interacts with a specified target while performing a desired action. The target is defined by a segmentation mask, and the action is described through a text prompt. Our key motivation is to incorporate target awareness into video generation, enabling actors to perform directed actions on designated objects. This enables video diffusion models to act as motion planners, producing plausible predictions of human–object interactions by leveraging the priors of large-scale video generative models. We build our target-aware model by extending a baseline model to incorporate the target mask as an additional input. To enforce target awareness, we introduce a special token that encodes the target's spatial information within the text prompt. We then fine-tune the model with our curated dataset using an additional cross-attention loss that aligns the cross-attention maps associated with this token with the input target mask. To further improve performance, we selectively apply this loss to the most semantically relevant attention regions and transformer blocks. Experimental results show that our target-aware model outperforms existing solutions in generating videos where actors interact accurately with the specified targets. We further demonstrate its efficacy in two downstream applications: zero-shot 3D HOI motion synthesis with physical plausibility and long-term video content creation.

Core Task Landscape

This paper addresses: **Target-Aware Video Generation with Human-Object Interactions**

A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

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

- **Controllable Video Generation with Interaction Modeling**
- **Text-Driven Human-Object Interaction Generation**
- **3D Human-Object Interaction Reconstruction and Tracking**
- **Motion Synthesis and Diffusion-Based HOI Generation**
- **Datasets, Benchmarks, and Evaluation for HOI Video**
- **Specialized Applications and Downstream Tasks**
- **Foundational Techniques and Representations**

Complete Taxonomy Tree

- Target-Aware Video Generation with Human-Object Interactions Survey Taxonomy
- Controllable Video Generation with Interaction Modeling
  - Spatial Control via Masks and Trajectories ★ (5 papers)
  - [0] Target-Aware Video Diffusion Models (Anon et al., 2026) View paper
  - [16] Mask2IV: Interaction-Centric Video Generation via Mask Trajectories (Li Gen, 2025) View paper
  - [24] Boximator: Generating Rich and Controllable Motions for Video Synthesis (Wang Jiawei, 2024) View paper
  - [27] VHOI: Controllable Video Generation of Human-Object Interactions from Sparse Trajectories via Motion Densification (Wanyue Zhang, 2025) View paper
  - [50] MATRIX: Mask Track Alignment for Interaction-aware Video Generation (Kim, 2025) View paper
  - Pose-Guided Human-Object Interaction Synthesis (3 papers)
  - [7] Multi-identity Human Image Animation with Structural Video Diffusion (Wang Zhenzhi, 2025) View paper
  - [12] AnchorCrafter: Animate CyberAnchors Saling Your Products via Human-Object Interacting Video Generation (Xu, 2024) View paper
  - Character Animation with Scene Interaction (2 papers)
  - [1] GenHSI: Controllable Generation of Human-Scene Interaction Videos (Li Ze-kun, 2025) View paper
  - [3] Mimo: Controllable character video synthesis with spatial decomposed modeling (Yifang Men, 2025) View paper
  - Multi-Concept and Multi-Identity Interaction Generation (2 papers)
  - [30] Interact-Custom: Customized Human Object Interaction Image Generation (Xu Zhu, 2025) View paper
  - [43] InterActHuman: Multi-Concept Human Animation with Layout-Aligned Audio Conditions (Wang Zhenzhi, 2025) View paper
- Text-Driven Human-Object Interaction Generation
  - Zero-Shot Text-to-Interaction Synthesis (3 papers)
  - [6] Zero-Shot Generation of Human-Object Interaction Videos (Nawhal Megha, 2022) View paper
  - [10] InterDreamer: Zero-Shot Text to 3D Dynamic Human-Object Interaction (Liang-Yan Gui, 2024) View paper
  - [18] InteractAnything: Zero-shot Human Object Interaction Synthesis via LLM Feedback and Object Affordance Parsing (Zhang Jin-lu, 2025) View paper
  - Text-Conditioned 3D HOI Synthesis (3 papers)

    - [11] Thor: Text to human-object interaction diffusion via relation intervention (Qianyang Wu, 2024) View paper
    - [22] Interfusion: Text-driven generation of 3d human-object interaction (Sisi Dai, 2024) View paper
    - [37] InteractMove: Text-Controlled Human-Object Interaction Generation in 3D Scenes with Movable Objects (Xinhao Cai, 2025) View paper
    - Language-Guided Motion Policy and Instruction Following (3 papers)
    - [13] Human-object interaction from human-level instructions (Wu Zhen, 2025) View paper
    - [17] Handsonvlm: Vision-language models for hand-object interaction prediction (Bao Chen, 2024) View paper
    - [29] Human-Object Interaction via Automatically Designed VLM-Guided Motion Policy (Shi Ye, 2025) View paper
- 3D Human-Object Interaction Reconstruction and Tracking
    - Occlusion-Aware 3D HOI Reconstruction (2 papers)
    - [2] Occlusion-Aware Temporally Consistent Amodal Completion for 3D Human-Object Interaction Reconstruction (Hyungjun Doh, 2025) View paper
    - [19] Visibility Aware Human-Object Interaction Tracking from Single RGB Camera (Xianghui Xie, 2023) View paper
    - Gaze-Guided and Fine-Grained Interaction Modeling (2 papers)
    - [38] Gaze-guided Hand-Object Interaction Synthesis: Dataset and Method (Tian Jie, 2024) View paper
    - [39] InteracTalker: Prompt-Based Human-Object Interaction with Co-Speech Gesture Generation (Sreehari Rajan, 2025) View paper
- Motion Synthesis and Diffusion-Based HOI Generation
    - Affordance and Physics-Based Motion Synthesis (3 papers)
    - [4] Task-Oriented Human-Object Interactions Generation with Implicit Neural Representations (Quanzhou Li, 2024) View paper
    - [8] PhysDreamer: Physics-Based Interaction with 3D Objects via Video Generation (Zhang TianYuan, 2024) View paper
    - [25] NIFTY: Neural Object Interaction Fields for Guided Human Motion Synthesis (Nilesh Kulkarni, 2023) View paper
    - Diffusion-Based 4D HOI Generation (5 papers)
    - [15] HunyuanVideo-HOMA: Generic Human-Object Interaction in Multimodal Driven Human Animation (Huang Ziyao, 2025) View paper
    - [20] AvatarGO: Zero-shot 4D Human-Object Interaction Generation and Animation (Cao, 2024) View paper
    - [26] AnchorHOI: Zero-shot Generation of 4D Human-Object Interaction via Anchor-based Prior Distillation (Sisi Dai, 2025) View paper
    - [28] ZeroHSI: Zero-Shot 4D Human-Scene Interaction by Video Generation (Li Hongjie, 2024) View paper
    - [36] HOMA: Towards Generic Human-Object Interaction in Multimodal Driven Human Animation with Weak Conditions (Ziyao Huang, 2025) View paper
    - Waypoint and Trajectory-Conditioned Motion Synthesis (2 papers)
    - [34] Object-Aware 4D Human Motion Generation (Gui, 2025) View paper
    - [42] Controllable Human-Object Interaction Synthesis (Jiaman Li, 2023) View paper
- Datasets, Benchmarks, and Evaluation for HOI Video (3 papers)
    - [31] HOIGen-1M: A Large-scale Dataset for Human-Object Interaction Video Generation (Liu Kun, 2025) View paper
    - [35] InterPose: Learning to Generate Human-Object Interactions from Large-Scale Web Videos (Zhang Yang-song, 2025) View paper
    - [49] Video Reality Test: Can AI-Generated ASMR Videos fool VLMs and Humans? (Jiaqi Wang, 2025) View paper
- Specialized Applications and Downstream Tasks
    - Robotics and Embodied AI Applications (2 papers)
    - [5] Gen2Act: Human Video Generation in Novel Scenarios enables Generalizable Robot Manipulation (Bharadhwaj, 2024) View paper
    - [23] Multimodal human-intent modeling for contextual robot-to-human handovers of arbitrary objects (Ren Hanwen, 2025) View paper
    - Video Editing and Content Manipulation (3 papers)
    - [32] LeviTor: 3D Trajectory Oriented Image-to-Video Synthesis (Hanlin Wang, 2024) View paper
    - [41] Inserting videos into videos (Lee Donghoon, 2019) View paper
    - [46] Understanding Object Dynamics for Interactive Image-to-Video Synthesis (Andreas Blattmann, 2021) View paper
    - Segmentation and Scene Understanding for HOI (3 papers)
    - [14] InterRVOS: Interaction-aware Referring Video Object Segmentation (Jin Woojeong, 2025) View paper
    - [44] Efficient Multi-Head Attention for Human-Object Interaction Recognition and Video Scene Graph Generation (Anfel Amirat, 2024) View paper
    - [48] Predicting Human-Object Interactions in Egocentric Videos (Manuel Benavent-Lledo, 2022) View paper
    - Hand-Object Interaction Video Generation (2 papers)
    - [33] Hierarchical Video Prediction using Relational Layouts for Human-Object Interactions (Navaneeth Bodla, 2021) View paper
    - [40] Open-world Hand-Object Interaction Video Generation Based on Structure and Contact-aware Representation (Haodong Yan, 2025) View paper
- Foundational Techniques and Representations
    - Object-Centric and Relational Representations (2 papers)
    - [9] Conditional Object-Centric Learning from Video (Thomas Kipf, 2021) View paper
    - [21] Semantic-Aware Human Object Interaction Image Generation (Zhu Xu, 2024) View paper
    - General Controllable Synthesis Frameworks (1 papers)
    - [45] Controllable image and video synthesis (Jiang, 2024) View paper

Narrative

Core task: target-aware video generation with human-object interactions. This field addresses the challenge of synthesizing realistic videos in which humans interact with specific objects in controllable, semantically meaningful ways. The taxonomy reveals several complementary research directions. Controllable Video Generation with Interaction Modeling focuses on spatial and temporal control mechanisms—such as masks, trajectories, and bounding boxes—that guide where and how interactions unfold (e.g., Boximator[24], Mask2IV[16]). Text-Driven Human-Object Interaction Generation emphasizes language-based conditioning to specify interaction semantics, while 3D Human-Object Interaction Reconstruction and Tracking tackles the geometric and pose estimation aspects needed for physically plausible contact. Motion Synthesis and Diffusion-Based HOI Generation explores generative models that produce diverse, natural human motions conditioned on object affordances. Meanwhile, Datasets, Benchmarks, and Evaluation provide the empirical infrastructure, Specialized Applications target downstream tasks like robotics or virtual anchors, and Foundational Techniques supply core representations and architectures.

Recent work has intensified around fine-grained spatial control and physically grounded interaction modeling. A dense cluster of methods leverages mask-based or trajectory-based guidance to steer diffusion models, balancing flexibility with precise object targeting (Target-Aware Video Diffusion[0], VHOI[27], MATRIX[50]). These approaches often grapple with trade-offs between open-ended creativity and strict adherence to user-specified constraints. Target-Aware Video Diffusion[0] sits squarely within the Spatial Control via Masks and Trajectories branch, emphasizing how explicit spatial cues can anchor object interactions during generation. Compared to neighbors like Boximator[24], which uses bounding-box annotations for layout control, or Mask2IV[16], which conditions on segmentation masks, the original paper appears to integrate target-specific priors more tightly into the diffusion process. This focus on target awareness distinguishes it from more general controllable generation schemes, positioning it as a specialized solution for scenarios demanding precise human-object coordination.

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. Mask2IV: Interaction-Centric Video Generation via Mask Trajectories

Authors: Li Gen, Zhao Bo, Gen Li, Yang Jianfei, Bo Zhao, et al. (9 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Generating interaction-centric videos, such as those depicting humans or robots interacting with objects, is crucial for embodied intelligence, as they provide rich and diverse visual priors for robot learning, manipulation policy training, and affordance reasoning. However, existing methods often struggle to model such complex and dynamic interactions. While recent studies show that masks can serve as effective control signals and enhance generation quality, obtaining dense and precise mask ann...

Relationship Analysis

Both papers belong to the 'Spatial Control via Masks and Trajectories' category, using mask-based spatial constraints to guide interaction dynamics in video generation. They overlap in their use of segmentation masks to specify target objects and their focus on generating human-object interaction videos with spatial control. However, the original paper (Target-Aware Video Diffusion Models) uses a single static mask with a special [TGT] token and cross-attention loss to achieve target awareness, while Mask2IV adopts a decoupled two-stage pipeline that first predicts motion trajectories for both actor and object, then generates videos conditioned on these predicted trajectories, eliminating the need for dense mask inputs during generation.

---

2. Boximator: Generating Rich and Controllable Motions for Video Synthesis

Authors: Wang Jiawei, Zhang, Yuchen, Jiawei Wang, Zou Jiaxin, et al. (15 authors total) | **Year/Venue**: 2024 • International Conference on Machine Learning | **URL**: View paper

Abstract

Generating rich and controllable motion is a pivotal challenge in video synthesis. We propose Boximator, a new approach for fine-grained motion control. Boximator introduces two constraint types: hard box and soft box. Users select objects in the conditional frame using hard boxes and then use either type of boxes to roughly or rigorously define the object's position, shape, or motion path in future frames. Boximator functions as a plug-in for existing video diffusion models. Its training proces...

Relationship Analysis

Both papers belong to the 'Spatial Control via Masks and Trajectories' category, using spatial constraints to guide video generation with human-object interactions. While the original paper (Target-Aware Video Diffusion Models) uses segmentation masks to specify target objects and employs cross-attention loss to align the [TGT] token with mask regions for interaction modeling, Boximator uses bounding boxes (hard and soft) as constraints and introduces a self-tracking technique to learn box-object correlations. The key difference is that the original paper focuses on target-aware interaction generation through mask-based spatial grounding in text conditioning, whereas Boximator emphasizes general motion control through box constraints as a plug-in module for existing video diffusion models.

---

3. VHOI: Controllable Video Generation of Human-Object Interactions from Sparse Trajectories via Motion Densification

Authors: Wanyue Zhang, Lin Geng Foo, Thabo Beeler, Rishabh Dabral, Christian Theobalt | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Synthesizing realistic human-object interactions (HOI) in video is challenging due to the complex, instance-specific interaction dynamics of both humans and objects. Incorporating controllability in video generation further adds to the complexity. Existing controllable video generation approaches face a trade-off: sparse controls like keypoint trajectories are easy to specify but lack instance-awareness, while dense signals such as optical flow, depths or 3D meshes are informative but costly to ...

Relationship Analysis

Both papers belong to the 'Spatial Control via Masks and Trajectories' category, using spatial constraints to guide human-object interactions in video generation. They overlap in their use of segmentation masks as control signals and their focus on generating realistic interaction dynamics. However, the original paper (TAVID) introduces a target-aware approach using a special [TGT] token with cross-attention loss to align attention maps with target masks, while the candidate paper (VHOI) employs a two-stage framework that densifies sparse trajectories into HOI mask sequences through a learned augmentor network, emphasizing part-level human semantics and trajectory-to-mask conversion.

---

4. MATRIX: Mask Track Alignment for Interaction-aware Video Generation

Authors: Kim, Seongchan, Siyoon Jin, Seongchan Kim, Lee Jae-Ho, et al. (16 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Video DiTs have advanced video generation, yet they still struggle to model multi-instance or subject-object interactions. This raises a key question: How do these models internally represent interactions? To answer this, we curate MATRIX-11K, a video dataset with interaction-aware captions and multi-instance mask tracks. Using this dataset, we conduct a systematic analysis that formalizes two perspectives of video DiTs: semantic grounding, via video-to-text attention, which evaluates whether no...

Relationship Analysis

Both papers belong to the 'Spatial Control via Masks and Trajectories' category, using mask-based spatial constraints to guide interaction dynamics in video generation. They overlap in their use of segmentation masks to specify target objects and their focus on enabling accurate human-object interactions through spatial grounding mechanisms. The key difference is that the original paper (Target-Aware Video Diffusion Models) introduces a special [TGT] token with cross-attention loss to align text prompts with target masks, while the candidate paper (MATRIX) proposes Semantic Grounding Alignment (SGA) and Semantic Propagation Alignment (SPA) losses applied to interaction-dominant layers, using multi-instance mask tracks to improve both grounding and temporal propagation of subject-object interactions.

Contributions Analysis

Overall novelty summary. The paper introduces a target-aware video diffusion model that generates videos where actors interact with specific objects defined by segmentation masks and text prompts. Within the taxonomy, it resides in the 'Spatial Control via Masks and Trajectories' leaf, which contains five papers total including this work. This leaf sits under 'Controllable Video Generation with Interaction Modeling', indicating a moderately populated research direction focused on explicit spatial guidance mechanisms. The sibling papers (Boximator, Mask2IV, VHOI, MATRIX) suggest an active but not overcrowded subfield exploring mask-based and trajectory-based control for interaction synthesis.

The taxonomy reveals neighboring research directions that contextualize this work's positioning. Adjacent leaves include 'Pose-Guided Human-Object Interaction Synthesis' (3 papers) emphasizing skeletal guidance, 'Character Animation with Scene Interaction' (2 papers) focusing on character-specific modeling, and 'Multi-Concept and Multi-Identity Interaction Generation' (2 papers) handling multiple subjects. The parent branch 'Controllable Video Generation with Interaction Modeling' excludes methods focused solely on 3D reconstruction or text-only control, clarifying that this work's mask-based spatial guidance distinguishes it from purely language-driven approaches in the sibling 'Text-Driven Human-Object Interaction Generation' branch.

Among 25 candidates examined across three contributions, no clearly refutable prior work was identified. The target-aware diffusion model with mask-based specification examined 10 candidates with zero refutations, the cross-attention loss with special token examined 5 candidates with zero refutations, and the curated dataset examined 10 candidates with zero refutations. This suggests that within the limited search scope, the specific combination of mask-based target specification, cross-attention alignment via special tokens, and selective loss application appears relatively unexplored. The sibling papers in the same taxonomy leaf employ related spatial control mechanisms but may differ in their attention-based grounding strategies or dataset curation approaches.

Based on the top-25 semantic matches examined, the work appears to occupy a distinct position within spatial control methods for interaction generation. The taxonomy structure indicates this is an active but not saturated research area, with the paper's specific technical approach—combining mask inputs, special token encoding, and selective cross-attention loss—not directly overlapped by the examined candidates. However, the limited search scope means potentially relevant work outside the top-25 matches or in adjacent subfields may exist but was not captured in this analysis.

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

Contribution 1: Target-aware video diffusion model with mask-based target specification

Description: The authors introduce a video diffusion model that accepts a segmentation mask to specify a target object and generates videos showing an actor performing text-prompted actions directed at that target. This enables explicit control over actor-target interactions without requiring dense motion annotations.

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. TokenMotion: Decoupled Motion Control via Token Disentanglement for Human-centric Video Generation

URL: View paper

Brief Assessment

TokenMotion[69] focuses on controlling camera motion and human poses in video generation, not on actor-target interactions specified by segmentation masks. The candidate addresses a different problem domain (human-centric motion control with camera trajectories) rather than enabling actors to interact with designated target objects through mask-based specification.

---

2. DanceTogether! Identity-Preserving Multi-Person Interactive Video Generation

URL: View paper

Brief Assessment

DanceTogether[64] focuses on multi-person identity preservation in interactive videos using pose-mask fusion, not on single-actor target-object interactions with mask-based target specification for action planning.

---

3. MGMAE: Motion Guided Masking for Video Masked Autoencoding

URL: View paper

Brief Assessment

MGMAE[72] focuses on masked autoencoding for video representation learning using motion-guided masking strategies, not on video generation or diffusion models with target specification for actor-target interactions.

---

4. Mask2IV: Interaction-Centric Video Generation via Mask Trajectories

URL: View paper

Brief Assessment

Mask2IV[16] focuses on a two-stage pipeline that first predicts motion trajectories then generates videos, rather than directly conditioning on masks during diffusion. The original paper's approach of using cross-attention loss to align mask inputs with special tokens differs fundamentally from Mask2IV's trajectory-based generation paradigm.

---

5. Intermask: 3d human interaction generation via collaborative masked modeling

URL: View paper

Brief Assessment

Intermask[71] focuses on 3D human-human interaction generation using collaborative masked modeling in discrete space, not video diffusion models for actor-target interactions with mask-based control.

---

6. MotionPro: A Precise Motion Controller for Image-to-Video Generation*

URL: View paper

Brief Assessment

MotionPro[66] focuses on precise motion control through region-wise trajectories and motion masks for general image animation, not on target-aware actor-object interaction synthesis using segmentation masks as spatial specifications for designated targets.

---

7. GenCompositor: Generative Video Compositing with Diffusion Transformer

URL: View paper

Brief Assessment

GenCompositor[65] focuses on video compositing by injecting foreground elements into target videos with user-controlled positioning, not on generating actor-target interactions from a single mask specification. The technical approaches and objectives differ fundamentally.

8. MAGREF: Masked Guidance for Any-Reference Video Generation

URL: View paper

Brief Assessment

MAGREF[67] addresses any-reference video generation with subject disentanglement for arbitrary reference subjects, which is a different problem from target-aware actor-target interaction synthesis using segmentation masks.

9. Diffusion Mask-Driven Visual-language Tracking

URL: View paper

Brief Assessment

Diffusion Mask-Driven Tracking[68] focuses on visual-language tracking tasks where masks are used to enhance tracking robustness, not on generating videos of actor-target interactions. The candidate addresses tracking existing objects in video sequences, while the original proposes generating new videos with controlled interactions.

10. Drag-A-Video: Non-rigid Video Editing with Point-based Interaction

URL: View paper

Brief Assessment

Drag-A-Video[70] focuses on interactive point-based video manipulation where users drag points to target locations, not on generating videos where actors perform text-prompted actions directed at mask-specified targets. The technical approaches differ fundamentally: Drag-A-Video uses point tracking and motion supervision for editing existing videos, while the original paper generates new videos with actor-target interactions using cross-attention alignment between mask inputs and special tokens.

Contribution 2: Cross-attention loss with special [TGT] token for spatial grounding

Description: The authors propose a training method that introduces a special [TGT] token in text prompts and applies a cross-attention loss to align the token's attention maps with the input target mask. This loss is selectively applied to semantically relevant attention regions and transformer blocks to enforce target awareness.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Personalization of Vision-language Models and the Multi-Concept Challenge

URL: View paper

Brief Assessment

Multi-Concept Challenge[62] focuses on personalization of vision-language models using learnable tokens to represent user-specific concepts in text-to-image generation, not on spatial grounding for video generation with target masks and cross-attention alignment.

2. Text-Image Alignment in Diffusion Models: The Role of Attention Sink

URL: View paper

Brief Assessment

Text-Image Alignment Attention[61] focuses on text-image semantic alignment in diffusion models through cross-attention analysis and self-attention guidance, not on spatial grounding of target objects via special tokens for video generation or human-object interaction tasks.

3. Exploring Cross-Attention Maps in Multi-modal Diffusion Transformers for Training-Free Semantic Segmentation

URL: View paper

Brief Assessment

Cross-Attention Maps Segmentation[59] focuses on training-free semantic segmentation using cross-attention maps from MM-DiT for image segmentation tasks, not on training video diffusion models with special tokens for spatial grounding in human-object interaction scenarios.

4. Improving global awareness of linkset predictions using Cross-Attentive Modulation tokens

URL: View paper

Brief Assessment

Cross-Attentive Modulation Tokens[63] focuses on graph neural networks for link prediction tasks, using cross-attention to gather global information for modulating node/edge embeddings. The original paper addresses spatial grounding in video diffusion models through cross-attention alignment with segmentation masks, which is a fundamentally different application domain and technical approach.

5. Not all diffusion model activations have been evaluated as discriminative features

URL: View paper

Brief Assessment

Diffusion Model Activations[60] focuses on selecting discriminative features from pre-trained diffusion models for tasks like semantic segmentation and correspondence, not on training methods with special tokens for spatial grounding in video generation.

Contribution 3: Curated dataset for target-aware video generation

Description: The authors construct a dataset of 1290 video clips from BEHAVE and Ego-Exo4D, where each clip shows an actor initially not interacting with a target and then engaging with it. Each video is annotated with target masks and text prompts describing the action.

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. CORE4D: A 4D Human-Object-Human Interaction Dataset for Collaborative Object REarrangement

URL: View paper

Brief Assessment

CORE4D[55] focuses on collaborative human-object-human rearrangement interactions with 4D motion capture data, while the original paper addresses single-actor target-aware video generation with segmentation masks and text prompts. The datasets serve fundamentally different purposes and interaction modalities.

2. Towards Human-like Virtual Beings: Simulating Human Behavior in 3D Scenes

URL: View paper

Brief Assessment

Human-like Virtual Beings[54] focuses on simulating high-level, long-horizon human behaviors in 3D scenes with hierarchical activity planning, not target-aware video generation with actor-object interaction annotations. Their BehaviorHub dataset contains linguistic goal-plan trees paired with 3D motion sequences, which differs fundamentally from the original paper's video clips with target masks and text prompts for video diffusion models.

---

3. Decoupled Generative Modeling for Human-Object Interaction Synthesis

URL: View paper

Brief Assessment

Decoupled Generative Modeling[57] focuses on human-object interaction synthesis using existing datasets (FullBodyManipulation and 3D-Future) for training their decoupled trajectory and action generators. The candidate does not present a dataset construction methodology for target-aware video generation with initial non-interaction frames transitioning to interaction frames, which is the core novelty of the original paper's dataset contribution.

---

4. InterRVOS: Interaction-aware Referring Video Object Segmentation

URL: View paper

Brief Assessment

InterRVOS[14] focuses on video object segmentation with actor-target interaction annotations for segmentation tasks, not video generation. The datasets serve fundamentally different purposes: InterRVOS[14] annotates segmentation masks for referring expressions, while the original paper constructs videos showing actors transitioning from non-interaction to interaction states for generative modeling.

---

5. NIFTY: Neural Object Interaction Fields for Guided Human Motion Synthesis

URL: View paper

Brief Assessment

NIFTY[25] focuses on 3D human motion synthesis with objects using anchor poses and synthetic data generation, not on video generation with target masks and text prompts for actor-object interactions as in the original paper.

---

6. Reasoning About Physical Interactions with Object-Oriented Prediction and Planning

URL: View paper

Brief Assessment

Object-Oriented Prediction Planning[51] focuses on learning object representations for physical reasoning tasks using synthetic block-stacking simulations, not on curating real-world video datasets with actor-object interaction annotations for target-aware video generation.

---

7. Object interaction-based surveillance video synopsis

URL: View paper

Brief Assessment

Object Interaction Surveillance[58] focuses on surveillance video synopsis and object interaction detection in security contexts, not on target-aware video generation with actor-object interaction annotations for training generative models.

---

8. Object Motion Guided Human Motion Synthesis

URL: View paper

Brief Assessment

Object Motion Guided Synthesis[52] focuses on full-body human motion synthesis for object manipulation using object motion as input, collecting a dataset of 3D object geometry and human-object interaction motions. This differs from the original paper's dataset of video clips with target masks and text prompts for target-aware video generation.

---

9. InterTrack: Tracking Human Object Interaction Without Object Templates

URL: View paper

Brief Assessment

InterTrack[53] focuses on tracking human-object interactions in videos using synthetic data for pose tracking, not on generating videos with target-aware interactions. The datasets serve fundamentally different purposes: InterTrack[53] creates synthetic videos for training tracking models, while the original paper curates real video clips for training video generation models.

---

10. G-HOP: Generative Hand-Object Prior for Interaction Reconstruction and Grasp Synthesis

URL: View paper

Brief Assessment

G-HOP[56] focuses on hand-object interaction datasets for 3D reconstruction and grasp synthesis, not target-aware video generation with actor-object interaction annotations as described in the original paper.

Appendix: Text Similarity Detection

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

References

- [0] Target-Aware Video Diffusion Models View paper
- [1] GenHSI: Controllable Generation of Human-Scene Interaction Videos View paper
- [2] Occlusion-Aware Temporally Consistent Amodal Completion for 3D Human-Object Interaction Reconstruction View paper
- [3] Mimo: Controllable character video synthesis with spatial decomposed modeling View paper
- [4] Task-Oriented Human-Object Interactions Generation with Implicit Neural Representations View paper
- [5] Gen2Act: Human Video Generation in Novel Scenarios enables Generalizable Robot Manipulation View paper
- [6] Zero-Shot Generation of Human-Object Interaction Videos View paper
- [7] Multi-identity Human Image Animation with Structural Video Diffusion View paper
- [8] PhysDreamer: Physics-Based Interaction with 3D Objects via Video Generation View paper

- [9] Conditional Object-Centric Learning from Video View paper
- [10] InterDreamer: Zero-Shot Text to 3D Dynamic Human-Object Interaction View paper
- [11] Thor: Text to human-object interaction diffusion via relation intervention View paper
- [12] AnchorCrafter: Animate CyberAnchors Saling Your Products via Human-Object Interacting Video Generation View paper
- [13] Human-object interaction from human-level instructions View paper
- [14] InterRVOS: Interaction-aware Referring Video Object Segmentation View paper
- [15] HunyuanVideo-HOMA: Generic Human-Object Interaction in Multimodal Driven Human Animation View paper
- [16] Mask2IV: Interaction-Centric Video Generation via Mask Trajectories View paper
- [17] Handsonvlm: Vision-language models for hand-object interaction prediction View paper
- [18] InteractAnything: Zero-shot Human Object Interaction Synthesis via LLM Feedback and Object Affordance Parsing View paper
- [19] Visibility Aware Human-Object Interaction Tracking from Single RGB Camera View paper
- [20] AvatarGO: Zero-shot 4D Human-Object Interaction Generation and Animation View paper
- [21] Semantic-Aware Human Object Interaction Image Generation View paper
- [22] Interfusion: Text-driven generation of 3d human-object interaction View paper
- [23] Multimodal human-intent modeling for contextual robot-to-human handovers of arbitrary objects View paper
- [24] Boximator: Generating Rich and Controllable Motions for Video Synthesis View paper
- [25] NIFTY: Neural Object Interaction Fields for Guided Human Motion Synthesis View paper
- [26] AnchorHOI: Zero-shot Generation of 4D Human-Object Interaction via Anchor-based Prior Distillation View paper
- [27] VHOI: Controllable Video Generation of Human-Object Interactions from Sparse Trajectories via Motion Densification View paper
- [28] ZeroHSI: Zero-Shot 4D Human-Scene Interaction by Video Generation View paper
- [29] Human-Object Interaction via Automatically Designed VLM-Guided Motion Policy View paper
- [30] Interact-Custom: Customized Human Object Interaction Image Generation View paper
- [31] HOIGen-1M: A Large-scale Dataset for Human-Object Interaction Video Generation View paper
- [32] LeviTor: 3D Trajectory Oriented Image-to-Video Synthesis View paper
- [33] Hierarchical Video Prediction using Relational Layouts for Human-Object Interactions View paper
- [34] Object-Aware 4D Human Motion Generation View paper
- [35] InterPose: Learning to Generate Human-Object Interactions from Large-Scale Web Videos View paper
- [36] HOMA: Towards Generic Human-Object Interaction in Multimodal Driven Human Animation with Weak Conditions View paper
- [37] InteractMove: Text-Controlled Human-Object Interaction Generation in 3D Scenes with Movable Objects View paper
- [38] Gaze-guided Hand-Object Interaction Synthesis: Dataset and Method View paper
- [39] InteracTalker: Prompt-Based Human-Object Interaction with Co-Speech Gesture Generation View paper
- [40] Open-world Hand-Object Interaction Video Generation Based on Structure and Contact-aware Representation View paper
- [41] Inserting videos into videos View paper
- [42] Controllable Human-Object Interaction Synthesis View paper
- [43] InterActHuman: Multi-Concept Human Animation with Layout-Aligned Audio Conditions View paper
- [44] Efficient Multi-Head Attention for Human-Object Interaction Recognition and Video Scene Graph Generation View paper
- [45] Controllable image and video synthesis View paper
- [46] Understanding Object Dynamics for Interactive Image-to-Video Synthesis View paper
- [47] AnchorCrafter: Animate Cyber-Anchors Selling Your Products via Human-Object Interacting Video Generation View paper
- [48] Predicting Human-Object Interactions in Egocentric Videos View paper
- [49] Video Reality Test: Can AI-Generated ASMR Videos fool VLMs and Humans? View paper
- [50] MATRIX: Mask Track Alignment for Interaction-aware Video Generation View paper
- [51] Reasoning About Physical Interactions with Object-Oriented Prediction and Planning View paper
- [52] Object Motion Guided Human Motion Synthesis View paper
- [53] InterTrack: Tracking Human Object Interaction Without Object Templates View paper
- [54] Towards Human-like Virtual Beings: Simulating Human Behavior in 3D Scenes View paper
- [55] CORE4D: A 4D Human-Object-Human Interaction Dataset for Collaborative Object REarrangement View paper
- [56] G-HOP: Generative Hand-Object Prior for Interaction Reconstruction and Grasp Synthesis View paper
- [57] Decoupled Generative Modeling for Human-Object Interaction Synthesis View paper
- [58] Object interaction-based surveillance video synopsis View paper
- [59] Exploring Cross-Attention Maps in Multi-modal Diffusion Transformers for Training-Free Semantic Segmentation View paper
- [60] Not all diffusion model activations have been evaluated as discriminative features View paper
- [61] Text-Image Alignment in Diffusion Models: The Role of Attention Sink View paper
- [62] Personalization of Vision-language Models and the Multi-Concept Challenge View paper
- [63] Improving global awareness of linkset predictions using Cross-Attentive Modulation tokens View paper
- [64] DanceTogether! Identity-Preserving Multi-Person Interactive Video Generation View paper
- [65] GenCompositor: Generative Video Compositing with Diffusion Transformer View paper
- [66] MotionPro: A Precise Motion Controller for Image-to-Video Generation* View paper
- [67] MAGREF: Masked Guidance for Any-Reference Video Generation View paper
- [68] Diffusion Mask-Driven Visual-language Tracking View paper
- [69] TokenMotion: Decoupled Motion Control via Token Disentanglement for Human-centric Video Generation View paper
- [70] Drag-A-Video: Non-rigid Video Editing with Point-based Interaction View paper
- [71] Intermask: 3d human interaction generation via collaborative masked modeling View paper
- [72] MGMAE: Motion Guided Masking for Video Masked Autoencoding View paper