Novelty Assessment Report

Paper: EgoTwin: Dreaming Body and View in First Person
PDF URL: https://openreview.net/pdf?id=QFJkvv3zMi
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

While exocentric video synthesis has achieved great progress, egocentric video generation remains largely underexplored, which requires modeling first-person view content along with camera motion patterns induced by the wearer's body movements. To bridge this gap, we introduce a novel task of joint egocentric video and human motion generation, characterized by two key challenges: 1) Viewpoint Alignment: the camera trajectory in the generated video must accurately align with the head trajectory derived from human motion; 2) Causal Interplay: the synthesized human motion must causally align with the observed visual dynamics across adjacent video frames. To address these challenges, we propose EgoTwin, a joint video-motion generation framework built on the diffusion transformer architecture. Specifically, EgoTwin introduces a head-centric motion representation that anchors the human motion to the head joint and incorporates a cybernetics-inspired interaction mechanism that explicitly captures the causal interplay between video and motion within attention operations. For comprehensive evaluation, we curate a large-scale real-world dataset of synchronized text-video-motion triplets and design novel metrics to assess video-motion consistency. Extensive experiments demonstrate the effectiveness of the EgoTwin framework. Qualitative results are available on our project page: https://egotwin.pages.dev/.

Core Task Landscape

This paper addresses: **Joint Egocentric Video and Human Motion Generation**

A total of **30 papers** were analyzed and organized into a taxonomy with **17 categories**.

Taxonomy Overview

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

- **Egocentric Motion Synthesis and Forecasting**
- **Egocentric Human-Object Interaction Modeling**
- **Egocentric Perception and Representation Learning**
- **Egocentric Benchmarks and Data Capture**
- **Video-Centric Generation and Simulation**
- **Applications and Unified Frameworks**

Complete Taxonomy Tree

- Joint Egocentric Video and Human Motion Generation Survey Taxonomy
- Egocentric Motion Synthesis and Forecasting
  - Joint Video-Motion Generation ★ (2 papers)
  - [0] EgoTwin: Dreaming Body and View in First Person (Anon et al., 2026) View paper
  - [1] Egogen: An egocentric synthetic data generator (Gen Li, 2024) View paper
  - Egocentric Pose Estimation and Forecasting (3 papers)
  - [14] 3d ego-pose estimation via imitation learning (Ye Yuan, 2018) View paper
  - [17] EgoCast: Forecasting Egocentric Human Pose in the Wild (MarÃa Escobar, 2024) View paper
  - [20] 3D Human pose estimation from egocentric inputs (Palomeque, 2024) View paper
  - Environment-Aware Motion Generation (2 papers)
  - [6] HMD2: Environment-Aware Motion Generation from Single Egocentric Head-Mounted Device (Guzov Vladimir, 2025) View paper
  - [16] UniEgoMotion: A Unified Model for Egocentric Motion Reconstruction, Forecasting, and Generation (Patel Chaitanya, 2025) View paper
  - Egocentric Pedestrian Motion Prediction (2 papers)
  - [12] Convolutional transformer network: Future pedestrian location in first-person videos using depth map and 3D pose (Kai Chen, 2023) View paper
  - [27] Egocentric pedestrian motion prediction by separately modeling body pose and position (Donghao Wu, 2020) View paper
- Egocentric Human-Object Interaction Modeling
  - Interaction Anticipation and Forecasting (3 papers)
  - [10] Forecasting human-object interaction: joint prediction of motor attention and actions in first person video (Miao Liu, 2020) View paper
  - [15] Ego-centric Predictive Model Conditioned on Hand Trajectories (Zhang Bin-jie, 2025) View paper
  - [30] MADiff: Motion-Aware Mamba Diffusion Models for Hand Trajectory Prediction on Egocentric Videos. (Ma, n.d.) View paper
  - Interaction Region Estimation (1 papers)
  - [11] Egochoir: Capturing 3d human-object interaction regions from egocentric views (Yang Cao, 2024) View paper
  - Scene Change Modeling from Interaction (1 papers)

- [7] Interaction replica: Tracking humanâ□□object interaction and scene changes from human motion (Vladimir Guzov, 2024) View paper
- Egocentric Perception and Representation Learning
  - Egocentric Activity Recognition (2 papers)
  - [24] PDEs on graphs for semi-supervised learning applied to first-person activity recognition in body-worn video (Hao Li, 2021) View paper
  - [26] STAC: Spatial-Temporal Attention on Compensation Information for Activity Recognition in FPV (Yue Zhang, 2021) View paper
  - Embodied Action Learning (3 papers)
  - [22] Learning embodied models of actions from first person video (Li Yin, 2017) View paper
  - [23] Force from motion: decoding physical sensation in a first person video (Hyun Soo Park, 2016) View paper
  - [28] Force from Motion: Decoding Control Force of Activity in a First-Person Video. (Hyun Soo Park, 2020) View paper
  - Gaze and Attention Analysis (1 papers)
  - [25] In the Eye of the Beholder: Gaze and Actions in First Person Video. (Yin Li, 2023) View paper
- Egocentric Benchmarks and Data Capture
  - Multi-Human Egocentric Benchmarks (1 papers)
  - [4] Ego-humans: An ego-centric 3d multi-human benchmark (Rawal Khirodkar, 2023) View paper
  - Egocentric Interaction Datasets (1 papers)
  - [3] Perceiving and Acting in First-Person: A Dataset and Benchmark for Egocentric Human-Object-Human Interactions (Xu Liang, 2025) View paper
- Video-Centric Generation and Simulation
  - Egocentric Avatar Animation (2 papers)
  - [19] EgoAnimate: Generating Human Animations from Egocentric top-down Views (Tanke, 2025) View paper
  - [29] EgoAnimate: Generating Human Animations from Egocentric top-down Views via Controllable Latent Diffusion Models (GK Turkoglu, n.d.) View paper
  - First-Person World Simulation (1 papers)
  - [9] The Matrix: Infinite-Horizon World Generation with Real-Time Moving Control (Feng Rui-li, 2024) View paper
- Applications and Unified Frameworks
  - Humanoid and Robotic Control (1 papers)
  - [2] Humanoid-VLA: Towards Universal Humanoid Control with Visual Integration (Ding, 2025) View paper
  - Motor Learning and Training Applications (4 papers)
  - [5] Rehearsal Reality: Exploring the Visualization of Experts' Mental Rehearsals from a First-Person Perspective to Support the Motor Learning of Novices (Yuto SUZUKI, 2025) View paper
  - [8] Improving viewing experiences of first-person shooter gameplays with automatically-generated motion effects (Gyeore Yun, 2021) View paper
  - [13] Multimodal Drumming Education Tool in Mixed Reality (James Pinkl, 2024) View paper
  - [18] Onebody: Remote posture guidance system using first person view in virtual environment (Vetere Frank, 2016) View paper
  - Unified Behavior Modeling (1 papers)
  - [21] Unified Simulation, Perception, and Generation of Human Behavior (Ye, 2022) View paper

Narrative

Core task: joint egocentric video and human motion generation. This field addresses the challenge of synthesizing both first-person visual content and corresponding full-body motion in a unified manner, bridging perception and embodied simulation. The taxonomy reveals several complementary directions: Egocentric Motion Synthesis and Forecasting focuses on predicting or generating body movements from egocentric cues, often leveraging head-mounted sensors or sparse observations; Egocentric Human-Object Interaction Modeling emphasizes the interplay between hands, objects, and scene context during manipulation tasks; Egocentric Perception and Representation Learning develops encoders and features tailored to the first-person viewpoint; Egocentric Benchmarks and Data Capture provides datasets and capture protocols (e.g., Interaction Replica[7], Ego Humans[4]) that ground these methods in real recordings; Video-Centric Generation and Simulation explores broader video synthesis techniques applicable to egocentric settings; and Applications and Unified Frameworks integrate these components for end-to-end systems in mixed reality, robotics, and content creation.

Within Motion Synthesis and Forecasting, a particularly active line of work targets joint video-motion generation, where models must produce coherent egocentric imagery alongside plausible body kinematics. EgoTwin[0] exemplifies this direction by jointly modeling visual and motion streams, ensuring temporal consistency between what the camera sees and how the body moves. Its closest neighbor, Egogen[1], similarly tackles egocentric video synthesis but may place different emphasis on conditioning modalities or diffusion architectures. Nearby efforts such as Rehearsal Reality[5] and Humanoid VLA[2] explore related themes—immersive simulation and embodied control—highlighting trade-offs between realism, interactivity, and computational cost. A key open question across these works is how to balance high-fidelity rendering with real-time motion forecasting, especially when integrating object interactions (e.g., Egocentric Human Object[3]) or adapting to diverse user behaviors captured in large-scale benchmarks.

Related Works in Same Category

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

1. Egogen: An egocentric synthetic data generator

Authors: Gen Li, Kaifeng Zhao, Siwei Zhang, Xiaozhong Lyu, Mihai Dusmanu, et al. (8 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â□¦ Our body movements are not solely a response to visual stimuli; they also change our â□¦ model requires jointly captured data of ground truth HMD trajectories and egocentric videos, which â□¦

Relationship Analysis

Both papers belong to the Joint Video-Motion Generation category, focusing on synthesizing egocentric video and human motion with viewpoint alignment. While EgoTwin generates egocentric video and motion jointly from text descriptions using a diffusion transformer with explicit causal interplay mechanisms, EgoGen focuses on creating a synthetic data generation system that uses reinforcement learning-based motion synthesis driven by egocentric visual sensing to produce training data for egocentric perception tasks. The key difference is that EgoTwin addresses joint generative modeling with bidirectional video-motion interaction, whereas EgoGen emphasizes scalable synthetic data generation with RL-based motion control for downstream task training.

Contributions Analysis

Overall novelty summary. The paper introduces a joint egocentric video and human motion generation task, positioning itself within the 'Joint Video-Motion Generation' leaf of the taxonomy. This leaf contains only two papers, including the original work, indicating a relatively sparse and emerging research direction. The task addresses simultaneous synthesis of first-person visual content and full-body kinematics with explicit viewpoint alignment and causal interplay constraints, distinguishing it from methods that generate motion or video independently.

The taxonomy reveals that neighboring leaves focus on complementary aspects: 'Egocentric Pose Estimation and Forecasting' predicts future poses from visual inputs but excludes simultaneous video synthesis; 'Environment-Aware Motion Generation' conditions motion on 3D scene context without generating corresponding imagery; and 'Egocentric Avatar Animation' produces human animations from top-down views but lacks the bidirectional video-motion coupling emphasized here. The paper's approach bridges perception-driven motion forecasting and video-centric generation, occupying a distinct niche at their intersection.

Among 21 candidates examined across three contributions, no clearly refutable prior work was identified. The joint task formulation examined 8 candidates with no refutations, suggesting novelty in the problem definition itself. The head-centric motion representation analyzed 3 candidates without overlap, indicating a potentially underexplored anchoring strategy. The EgoTwin framework reviewed 10 candidates with no refutations, though the limited search scope means comprehensive diffusion transformer architectures for this dual-modality setting may exist beyond the examined set.

Based on the top-21 semantic matches and taxonomy structure, the work appears to occupy a sparse research area with few direct competitors. The analysis covers contributions at the task, representation, and framework levels but does not exhaustively survey all diffusion-based video generation or motion synthesis methods. The sibling paper in the same leaf and nearby taxonomy branches provide context, yet the limited candidate pool suggests caution in claiming definitively novel territory without broader literature coverage.

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

Contribution 1: Joint egocentric video and human motion generation task

Description: The authors define a new task that requires generating synchronized egocentric video and human motion sequences from text descriptions, an initial pose, and an initial observation. This task explicitly models the tight coupling between camera motion and body movement in first-person perspectives, addressing viewpoint alignment and causal interplay challenges.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Human Motion Aware Text-to-Video Generation with Explicit Camera Control

URL: View paper

Brief Assessment

Motion Camera Control[36] focuses on generating exocentric (third-person) videos with human motion using skeleton projections, not egocentric (first-person) video generation with viewpoint alignment and causal interplay challenges.

---

2. Decomposing Text into Motion and Appearance for Training-Free Human Video Generation

URL: View paper

Brief Assessment

Decomposing Motion Appearance[38] focuses on text-to-human video generation by decomposing text into motion and appearance components for third-person videos, not on joint egocentric video and human motion generation with viewpoint alignment challenges.

---

3. Dancing Avatar: Pose and Text-Guided Human Motion Videos Synthesis with Image Diffusion Model

URL: View paper

Brief Assessment

Dancing Avatar[33] focuses on generating human motion videos from text and pose sequences using a T2I diffusion model in an autoregressive manner, not on joint egocentric video-motion generation with viewpoint alignment and causal interplay as defined in the original paper.

---

4. Exocentric-to-egocentric video generation

URL: View paper

Brief Assessment

Exocentric to Egocentric[31] focuses on generating egocentric videos from exocentric (third-person) multi-view videos, not on jointly generating egocentric video with synchronized human motion from text descriptions as in the original paper.

---

5. Enhancing Human-Computer Interaction Through Decoupling Motion and Camera Control in Human-Centric Video Generation

URL: View paper

Brief Assessment

Decoupling Motion Camera[35] focuses on decoupling motion and camera control within text-to-video generation for precise adjustments, not on the joint generation of synchronized egocentric video and human motion sequences with viewpoint alignment and causal interplay as defined in the original paper.

---

6. Text-Based Video Generation With Human Motion and Controllable Camera

URL: View paper

Brief Assessment

Text Motion Camera[37] focuses on exocentric (third-person) video generation with controllable camera angles using pre-generated 3D human motion, not joint egocentric video-motion generation with viewpoint alignment and causal interplay as in the original paper.

---

7. Generating Human Motion Videos using a Cascaded Text-to-Video Framework

URL: View paper

Brief Assessment

Cascaded Motion Videos[34] focuses on generating human motion videos using a cascaded text-to-video framework that bridges text-to-motion models with video diffusion models, but does not address the specific egocentric video generation task with viewpoint alignment and causal interplay challenges that the original paper tackles.

---

8. Open-set synthesis for free-viewpoint human body reenactment of novel poses

URL: View paper

Brief Assessment

Open Set Synthesis[32] focuses on multi-view pose synthesis for novel poses with viewpoint consistency, not on joint egocentric video and human motion generation from text descriptions.

Contribution 2: Head-centric motion representation

Description: The authors propose a novel motion representation that explicitly exposes head joint pose and velocity, replacing the conventional root-centric representation. This reformulation facilitates accurate alignment between egocentric camera trajectories and head motion, which is critical for viewpoint consistency.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Mocap Everyone Everywhere: Lightweight Motion Capture with Smartwatches and a Head-Mounted Camera

URL: View paper

Brief Assessment

Mocap Everyone Everywhere[49] focuses on motion capture from smartwatches and head-mounted cameras, using head poses as input signals for motion estimation. The original paper proposes a head-centric motion representation for joint video-motion generation to align camera viewpoints with head motion in egocentric video synthesis, which is a fundamentally different application domain and technical approach.

---

2. Large eyeâ□□head gaze shifts measured with a wearable eye tracker and an industrial camera

URL: View paper

Brief Assessment

Large Eye Head[50] focuses on measuring eye-head gaze shifts using wearable eye trackers and industrial cameras for tracking head orientation, not on motion representation for synthesis or generation tasks.

---

3. ECHO: Ego-Centric modeling of Human-Object interactions

URL: View paper

Brief Assessment

ECHO[51] uses a head-centric coordinate system for modeling human-object interactions from sparse 3-point tracking, but does not propose a motion representation that explicitly exposes head joint pose and velocity for aligning camera trajectories with head motion in video synthesis. The original paper's contribution focuses on reformulating motion representation to facilitate viewpoint alignment between egocentric video and head motion in joint video-motion generation, which is a different task than ECHO[51]'s interaction reconstruction from sparse sensors.

Contribution 3: EgoTwin diffusion transformer framework with interaction mechanism

Description: The authors develop a triple-branch diffusion transformer architecture with modality-specific branches for text, video, and motion. The framework employs asynchronous diffusion and a structured attention mask inspired by cybernetic observation-action loops to model the causal interplay between visual observations and human actions.

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. Edit-Your-Motion: Space-Time Diffusion Decoupling Learning for Video Motion Editing

URL: View paper

Brief Assessment

Edit Your Motion[44] focuses on video motion editing for human pose manipulation in existing videos, not joint generation of egocentric video and human motion with causal interplay modeling between modalities.

---

2. Improved video vae for latent video diffusion model

URL: View paper

Brief Assessment

Improved Video VAE[40] focuses on video VAE architecture for compression/reconstruction in latent diffusion models, not on joint video-motion generation or causal interplay modeling between visual observations and human actions.

---

3. From slow bidirectional to fast autoregressive video diffusion models

URL: View paper

Brief Assessment

Fast Autoregressive Video[39] focuses on distilling bidirectional video diffusion models into autoregressive generators for streaming generation, not on joint video-motion generation with causal interplay modeling between modalities as in EgoTwin.

---

4. Video2Act: A Dual-System Video Diffusion Policy with Robotic Spatio-Motional Modeling

URL: View paper

Brief Assessment

Video2Act[48] focuses on robotic manipulation using video diffusion models for spatial and motion feature extraction, not on joint egocentric video-motion generation with viewpoint alignment and causal interplay modeling as in the original paper.

---

5. Tora2: Motion and Appearance Customized Diffusion Transformer for Multi-Entity Video Generation

URL: View paper

Brief Assessment

Tora2[45] focuses on multi-entity video generation with motion and appearance customization, not on joint egocentric video-motion generation with causal interplay modeling between visual observations and human actions.

---

6. Long-Term TalkingFace Generation via Motion-Prior Conditional Diffusion Model

URL: View paper

Brief Assessment

Long Term TalkingFace[46] focuses on talking face video generation with audio-driven facial motion, not egocentric video-motion generation with causal observation-action loops between visual observations and human body movements.

---

7. The Quest for Generalizable Motion Generation: Data, Model, and Evaluation

URL: View paper

Brief Assessment

Generalizable Motion Generation[47] focuses on text-to-motion generation using flow-matching diffusion transformers with video priors for semantic diversity, not on joint egocentric video-motion generation with causal interplay modeling between modalities as in the original paper.

---

8. Co-Speech Gesture Video Generation via Motion-Decoupled Diffusion Model

URL: View paper

Brief Assessment

Co Speech Gesture[43] focuses on co-speech gesture video generation using TPS-based motion features and skeleton-free approaches, not on egocentric video-motion generation with causal observation-action loops.

---

9. LMP: Leveraging Motion Prior in Zero-Shot Video Generation with Diffusion Transformer

URL: View paper

Brief Assessment

LMP[42] focuses on motion transfer from reference videos to control subject motion in text-to-video and image-to-video generation, not on joint egocentric video-motion generation with causal interplay modeling between modalities.

---

10. OmniTalker: One-shot Real-time Text-Driven Talking Audio-Video Generation With Multimodal Style Mimicking

URL: View paper

Brief Assessment

OmniTalker[41] focuses on audio-video generation for talking heads with a dual-branch architecture for speech and facial synthesis, not egocentric video-motion generation with causal interplay modeling.

Appendix: Text Similarity Detection

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

References

- [0] EgoTwin: Dreaming Body and View in First Person View paper
- [1] Egogen: An egocentric synthetic data generator View paper
- [2] Humanoid-VLA: Towards Universal Humanoid Control with Visual Integration View paper
- [3] Perceiving and Acting in First-Person: A Dataset and Benchmark for Egocentric Human-Object-Human Interactions View paper
- [4] Ego-humans: An ego-centric 3d multi-human benchmark View paper
- [5] Rehearsal Reality: Exploring the Visualization of Experts' Mental Rehearsals from a First-Person Perspective to Support the Motor Learning of Novices View paper
- [6] HMD2: Environment-Aware Motion Generation from Single Egocentric Head-Mounted Device View paper
- [7] Interaction replica: Tracking humanâ□□object interaction and scene changes from human motion View paper
- [8] Improving viewing experiences of first-person shooter gameplays with automatically-generated motion effects View paper
- [9] The Matrix: Infinite-Horizon World Generation with Real-Time Moving Control View paper
- [10] Forecasting human-object interaction: joint prediction of motor attention and actions in first person video View paper
- [11] Egochoir: Capturing 3d human-object interaction regions from egocentric views View paper
- [12] Convolutional transformer network: Future pedestrian location in first-person videos using depth map and 3D pose View paper
- [13] Multimodal Drumming Education Tool in Mixed Reality View paper
- [14] 3d ego-pose estimation via imitation learning View paper
- [15] Ego-centric Predictive Model Conditioned on Hand Trajectories View paper
- [16] UniEgoMotion: A Unified Model for Egocentric Motion Reconstruction, Forecasting, and Generation View paper
- [17] EgoCast: Forecasting Egocentric Human Pose in the Wild View paper
- [18] Onebody: Remote posture guidance system using first person view in virtual environment View paper
- [19] EgoAnimate: Generating Human Animations from Egocentric top-down Views View paper
- [20] 3D Human pose estimation from egocentric inputs View paper
- [21] Unified Simulation, Perception, and Generation of Human Behavior View paper
- [22] Learning embodied models of actions from first person video View paper
- [23] Force from motion: decoding physical sensation in a first person video View paper
- [24] PDEs on graphs for semi-supervised learning applied to first-person activity recognition in body-worn video View paper
- [25] In the Eye of the Beholder: Gaze and Actions in First Person Video. View paper
- [26] STAC: Spatial-Temporal Attention on Compensation Information for Activity Recognition in FPV View paper
- [27] Egocentric pedestrian motion prediction by separately modeling body pose and position View paper
- [28] Force from Motion: Decoding Control Force of Activity in a First-Person Video. View paper
- [29] EgoAnimate: Generating Human Animations from Egocentric top-down Views via Controllable Latent Diffusion Models View paper
- [30] MADiff: Motion-Aware Mamba Diffusion Models for Hand Trajectory Prediction on Egocentric Videos. View paper
- [31] Exocentric-to-egocentric video generation View paper
- [32] Open-set synthesis for free-viewpoint human body reenactment of novel poses View paper
- [33] Dancing Avatar: Pose and Text-Guided Human Motion Videos Synthesis with Image Diffusion Model View paper
- [34] Generating Human Motion Videos using a Cascaded Text-to-Video Framework View paper
- [35] Enhancing Human-Computer Interaction Through Decoupling Motion and Camera Control in Human-Centric Video Generation View paper
- [36] Human Motion Aware Text-to-Video Generation with Explicit Camera Control View paper
- [37] Text-Based Video Generation With Human Motion and Controllable Camera View paper
- [38] Decomposing Text into Motion and Appearance for Training-Free Human Video Generation View paper
- [39] From slow bidirectional to fast autoregressive video diffusion models View paper
- [40] Improved video vae for latent video diffusion model View paper

- [41] OmniTalker: One-shot Real-time Text-Driven Talking Audio-Video Generation With Multimodal Style Mimicking View paper
- [42] LMP: Leveraging Motion Prior in Zero-Shot Video Generation with Diffusion Transformer View paper
- [43] Co-Speech Gesture Video Generation via Motion-Decoupled Diffusion Model View paper
- [44] Edit-Your-Motion: Space-Time Diffusion Decoupling Learning for Video Motion Editing View paper
- [45] Tora2: Motion and Appearance Customized Diffusion Transformer for Multi-Entity Video Generation View paper
- [46] Long-Term TalkingFace Generation via Motion-Prior Conditional Diffusion Model View paper
- [47] The Quest for Generalizable Motion Generation: Data, Model, and Evaluation View paper
- [48] Video2Act: A Dual-System Video Diffusion Policy with Robotic Spatio-Motional Modeling View paper
- [49] Mocap Everyone Everywhere: Lightweight Motion Capture with Smartwatches and a Head-Mounted Camera View paper
- [50] Large eyeâ□□head gaze shifts measured with a wearable eye tracker and an industrial camera View paper
- [51] ECHO: Ego-Centric modeling of Human-Object interactions View paper