Novelty Assessment Report

Paper: Latent Particle World Models: Self-supervised Object-centric Stochastic Dynamics Modeling
PDF URL: https://openreview.net/pdf?id=lTaPtGiUUc
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

We introduce Latent Particle World Model (LPWM), a self-supervised object-centric world model scaled to real-world multi-object datasets and applicable in decision-making. LPWM autonomously discovers keypoints, bounding boxes, and object masks directly from video data, enabling it to learn rich scene decompositions without supervision. Our architecture is trained end-to-end purely from videos and supports flexible conditioning on actions, language, and image goals. LPWM models stochastic particle dynamics via a novel latent action module and achieves state-of-the-art results on diverse real-world and synthetic datasets. Beyond stochastic video modeling, LPWM is readily applicable to decision-making, including goal-conditioned imitation learning, as we demonstrate in the paper. Code, and pre-trained models will be made publicly available. Video rollouts are available: https://sites.google.com/view/lpwm

Core Task Landscape

This paper addresses: **object-centric stochastic video prediction and world modeling**

A total of **50 papers** were analyzed and organized into a taxonomy with **27 categories**.

Taxonomy Overview

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

- **Object-Centric Representation Learning and Decomposition**
- **Object-Centric Video Prediction and Dynamics Modeling**
- **Language-Conditioned and Goal-Conditioned Object-Centric Models**
- **Holistic Video Generation and World Models**
- **Domain-Specific World Models and Applications**
- **Physics-Grounded and Interpretable World Models**
- **Structured and Compositional World Model Frameworks**
- **World Models for Reinforcement Learning and Planning**
- **Specialized Prediction Tasks and Modalities**
- **Text-Based and Symbolic World Models**
- ... and 1 more categories

Complete Taxonomy Tree

- object-centric stochastic video prediction and world modeling Survey Taxonomy
- Object-Centric Representation Learning and Decomposition
  - Slot-Based Object Discovery and Generative Modeling (4 papers)
  - [1] Object-centric Video Prediction without Annotation (Karl Schmeckpeper, 2021) View paper
  - [6] SlotDiffusion: Object-Centric Generative Modeling with Diffusion Models (Wu, 2023) View paper
  - [22] Advances in object-centric learning methods towards real world applications (Jindong, 2025) View paper
  - [25] Learning Object-Centric Representations Based on Slots in Real World Scenarios (Akan, 2025) View paper
  - Particle-Based and Graph-Based Object Modeling ★ (3 papers)
  - [0] Latent Particle World Models: Self-supervised Object-centric Stochastic Dynamics Modeling (Anon et al., 2026) View paper
  - [12] GraphMimic: Graph-to-Graphs Generative Modeling from Videos for Policy Learning (Guang-yan Chen, 2025) View paper
  - [32] OCK: Unsupervised Dynamic Video Prediction with Object-Centric Kinematics (Yeon-Ji Song, 2024) View paper
  - Scalable Object-Centric Architectures (1 papers)
  - [9] Scalor: Generative world models with scalable object representations (Jindong Jiang, 2019) View paper
- Object-Centric Video Prediction and Dynamics Modeling
  - Transformer-Based Object-Centric Prediction (3 papers)
  - [21] Slotformer: Unsupervised visual dynamics simulation with object-centric models (Wu, 2022) View paper
  - [23] Generative video transformer: Can objects be the words? (WU Yi-fu, 2021) View paper
  - [48] Patch-based Object-centric Transformers for Efficient Video Generation (Yan, 2022) View paper
  - Recurrent and State-Space Object Dynamics (3 papers)
  - [16] Deep variational Luenberger-type observer with dynamic objects channel-attention for stochastic video prediction (Dong Wang, 2023) View paper
  - [20] Stochastic prediction of multi-agent interactions from partial observations (Sun Chen, 2019) View paper
  - [39] A Symmetric and Object-Centric World Model for Stochastic Environments (Patrick Emami, 2020) View paper
  - Multi-View and Compositional Object Prediction (2 papers)
  - [13] Time-Conditioned Generative Modeling of Object-Centric Representations for Video Decomposition and Prediction (Gao, 2023) View paper

    - [34] Compositional video prediction (Ye Yufei, 2019) View paper
    - Physics-Grounded Object Dynamics (2 papers)
    - [31] Interaction Aware Relational Representations for Video Prediction (Rei Tamaru, 2021) View paper
    - [50] Learning Physical Dynamics for Object-centric Visual Prediction (Xu Huilin, 2024) View paper
- Language-Conditioned and Goal-Conditioned Object-Centric Models
    - Language-Guided Object-Centric Prediction (2 papers)
    - [14] Object-Centric World Model for Language-Guided Manipulation (Cha, 2025) View paper
    - [24] TextOCVP: Object-Centric Video Prediction with Language Guidance (Villar-Corrales, 2025) View paper
    - Goal-Conditioned Object-Centric World Models (2 papers)
    - [7] PlaySlot: Learning Inverse Latent Dynamics for Controllable Object-Centric Video Prediction and Planning (Villar-Corrales, 2025) View paper
    - [35] SOLD: Slot Object-Centric Latent Dynamics Models for Relational Manipulation Learning from Pixels (Mosbach, 2024) View paper
- Holistic Video Generation and World Models
    - Diffusion-Based Video Generation (2 papers)
    - [18] Wcdt: World-Centric Diffusion Transformer for Traffic Scene Generation (Chen Yang, 2024) View paper
    - [40] SAMPO:Scale-wise Autoregression with Motion PrOmpt for generative world models (Wang sen, 2025) View paper
    - Autoregressive and Masked Token World Models (3 papers)
    - [46] Generative World Modelling for Humanoids: 1X World Model Challenge Technical Report (Mereu, 2025) View paper
    - [47] WorldDreamer: Towards General World Models for Video Generation via Predicting Masked Tokens (Wang Xiaofeng, 2024) View paper
    - [49] Pandora: Towards General World Model with Natural Language Actions and Video States (Xiang Jiannan, 2024) View paper
    - Vision Foundation Model Feature Forecasting (1 papers)
    - [37] VFMF: World Modeling by Forecasting Vision Foundation Model Features (Gabrijel Boduljak, 2025) View paper
    - Hybrid Autoregressive-Causal World Models (1 papers)
    - [38] DyMoDreamer: World Modeling with Dynamic Modulation (Zhang Boxuan, 2025) View paper
- Domain-Specific World Models and Applications
    - Robotic Manipulation and Control (3 papers)
    - [2] Robot Learning from a Physical World Model (Jiageng Mao, 2025) View paper
    - [3] Spatial policy: Guiding visuomotor robotic manipulation with spatial-aware modeling and reasoning (Liu Yi-jun, 2025) View paper
    - [43] WorldGym: World Model as An Environment for Policy Evaluation (Sharma, 2025) View paper
    - Autonomous Driving and Navigation (4 papers)
    - [4] DriVerse: Navigation World Model for Driving Simulation via Multimodal Trajectory Prompting and Motion Alignment (Li XiaoFan, 2025) View paper
    - [8] UniUGP: Unifying Understanding, Generation, and Planing For End-to-end Autonomous Driving (Hao Lu, 2025) View paper
    - [26] Self-supervised multi-future occupancy forecasting for autonomous driving (Bernard Lange, 2024) View paper
    - [36] Seeing the Future, Perceiving the Future: A Unified Driving World Model for Future Generation and Perception (Liang, 2025) View paper
    - Egocentric and Embodied Vision (1 papers)
    - [15] GEM: A Generalizable Ego-Vision Multimodal World Model for Fine-Grained Ego-Motion, Object Dynamics, and Scene Composition Control (Mariam Hassan, 2024) View paper
    - Remote Sensing and Environmental Monitoring (1 papers)
    - [19] Scalable Multi-Temporal Remote Sensing Change Data Generation via Simulating Stochastic Change Process (Zhuo Zheng, 2023) View paper
    - Telecommunications and Network Modeling (1 papers)
    - [17] Agentic World Modeling for 6G: Near-Real-Time Generative State-Space Reasoning (Rezazadeh, 2025) View paper
    - Dynamic Occupancy and Spatial Forecasting (1 papers)
    - [11] SCOPE: Stochastic Cartographic Occupancy Prediction Engine for Uncertainty-Aware Dynamic Navigation (Zhanteng Xie, 2024) View paper
- Physics-Grounded and Interpretable World Models (2 papers)
    - [5] Physics-Grounded Motion Forecasting via Equation Discovery for Trajectory-Guided Image-to-Video Generation (Feng Tao, 2025) View paper
    - [10] Physgen: Rigid-body physics-grounded image-to-video generation (Shaowei Liu, 2024) View paper
- Structured and Compositional World Model Frameworks (1 papers)
    - [45] Natural Building Blocks for Structured World Models: Theory, Evidence, and Scaling (Da Costa, 2025) View paper
- World Models for Reinforcement Learning and Planning (2 papers)
    - [28] Object-Centric Dreamer (Leonid Ugadiarov, 2025) View paper
    - [42] AXIOM: Learning to Play Games in Minutes with Expanding Object-Centric Models (Heins, 2025) View paper
- Specialized Prediction Tasks and Modalities
    - Multi-Agent Interaction Prediction (1 papers)
    - [33] GenTrack: A New Generation of Multi-Object Tracking (Kraft Dirk, 2025) View paper
    - Human Activity and Temporal Transformation Prediction (2 papers)
    - [29] Predicting human activities using stochastic grammar (Siyuan Qi, 2017) View paper
    - [30] Learning temporal transformations from time-lapse videos (Zhou YiPin, 2016) View paper
    - Probabilistic Object Tracking and Indexing (1 papers)
    - [44] Effectively Indexing Uncertain Moving Objects for Predictive Queries (Meihui Zhang, 2009) View paper
- Text-Based and Symbolic World Models (1 papers)
    - [27] PLM-based World Models for Text-based Games (Hwang, 2022) View paper
- 3D and Geometric World Modeling (1 papers)
    - [41] 3D Video Models through Point Tracking, Reconstructing and Forecasting (Chu, 2025) View paper

Narrative

Core task: object-centric stochastic video prediction and world modeling. This field aims to learn structured representations that decompose visual scenes into distinct entities and predict their future states under uncertainty. The taxonomy reveals a rich landscape organized around several complementary themes. One major branch focuses on object-centric representation learning and decomposition, where methods develop techniques to discover and track entities using slots, particles, or graph-based structures (e.g., Scalor[9], SlotDiffusion[6]). Another branch emphasizes object-centric video prediction and dynamics modeling, building forward models that leverage these decomposed representations (e.g., Object-centric Video Prediction[1], Slotformer[21]). Additional branches address language-conditioned and goal-conditioned models, holistic video generation and world models for broader scene synthesis, domain-specific applications such as autonomous driving (DriVerse[4]) or robotics (Robot Physical World Model[2]), physics-grounded and interpretable approaches (Physgen[10], Physics-Grounded Motion Forecasting[5]), structured and compositional frameworks, world models tailored for reinforcement learning and planning, specialized prediction tasks, text-based and symbolic representations, and 3D geometric modeling.

Within this landscape, a particularly active line of work explores particle-based and graph-based object modeling, which represents entities as sets of interacting particles or nodes rather than fixed-size slot vectors. Latent Particle World Models[0] exemplifies this direction by using particle representations to capture fine-grained spatial structure and relational dynamics in a stochastic setting. This approach contrasts with slot-based methods like SlotDiffusion[6] or Slotformer[21], which typically employ a fixed number of abstract feature vectors, and aligns more closely with graph-structured models such as GraphMimic[12] and OCK[32], which emphasize explicit relational reasoning and flexible entity counts. The particle-based paradigm offers potential advantages in handling variable numbers of objects and modeling complex interactions, though it also raises questions about scalability and the trade-off between expressiveness and computational efficiency. Situating Latent Particle World Models[0] in this context, it occupies a niche that bridges fine-grained spatial decomposition with stochastic dynamics, offering a complementary perspective to both holistic generation approaches and more abstract slot-based frameworks.

Related Works in Same Category

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

1. GraphMimic: Graph-to-Graphs Generative Modeling from Videos for Policy Learning

Authors: Guang-yan Chen, Te Cui, Meiling Wang, Chengcai Yang, Mengxiao Hu, et al. (12 authors total) | **Year/Venue**: 2025 • Computer Vision and Pattern Recognition | **URL**: View paper

Abstract

Learning from demonstration is a powerful method for robotic skill acquisition. However, the significant expense of collecting such action-labeled robot data presents a major bottleneck. Video data, a rich data source encompassing diverse behavioral and physical knowledge, emerges as a promising alternative. In this paper, we present GraphMimic, a novel paradigm that leverages video data via graph-to-graphs generative modeling, which pre-trains models to generate future graphs conditioned on the...

Relationship Analysis

Both papers belong to the Particle-Based and Graph-Based Object Modeling category, representing objects as structured entities (particles or graph nodes) to model dynamics and interactions. GraphMimic overlaps with LPWM in using structured object-centric representations for learning dynamics from video, but differs fundamentally in its approach: GraphMimic constructs explicit object-action graphs from videos for policy learning via graph-to-graph generation, while LPWM learns implicit particle-based representations with latent actions through end-to-end VAE training for stochastic video prediction and world modeling.

---

2. OCK: Unsupervised Dynamic Video Prediction with Object-Centric Kinematics

Authors: Yeon-Ji Song, Jaein Kim, Suhyung Choi, Jin-Hwa Kim, Byoung-Tak Zhang | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Human perception involves decomposing complex multi-object scenes into time-static object appearance (i.e., size, shape, color) and time-varying object motion (i.e., position, velocity, acceleration). For machines to achieve human-like intelligence in real-world interactions, understanding these physical properties of objects is essential, forming the foundation for dynamic video prediction. While recent advancements in object-centric transformers have demonstrated potential in video prediction,...

Relationship Analysis

Both papers belong to the Particle-Based and Graph-Based Object Modeling category, representing objects through structured representations to model dynamics. OCK and LPWM overlap in their use of object-centric representations for video prediction, with both employing transformer-based architectures and learning decomposed scene representations in an unsupervised manner. However, LPWM uses latent particle representations with explicit keypoints and per-particle latent actions for stochastic dynamics modeling, while OCK focuses on explicit object kinematics (position, velocity, acceleration) integrated with object slots, emphasizing motion dynamics as a separate attribute from appearance.

Contributions Analysis

Overall novelty summary. The paper introduces a self-supervised object-centric world model using particle-based representations to discover keypoints, bounding boxes, and masks from video data. It sits within the 'Particle-Based and Graph-Based Object Modeling' leaf of the taxonomy, which contains only three papers total including this work. This represents a relatively sparse research direction compared to neighboring areas like 'Slot-Based Object Discovery' (four papers) or 'Transformer-Based Object-Centric Prediction' (three papers), suggesting the particle-based paradigm remains less explored than slot-based alternatives despite its potential for handling variable object counts and fine-grained spatial structure.

The taxonomy reveals substantial activity in adjacent areas. The parent branch 'Object-Centric Representation Learning' includes slot-based methods that use fixed-size feature vectors rather than flexible particle sets. Nearby branches address temporal dynamics through transformers or recurrent architectures, while 'Language-Conditioned and Goal-Conditioned Models' explores conditioning mechanisms similar to those claimed here. The paper's positioning bridges representation learning (particle discovery) with dynamics modeling (stochastic prediction) and decision-making applications, connecting multiple taxonomy branches. This cross-cutting nature distinguishes it from works focused solely on representation or prediction.

Among 26 candidates examined across three contributions, no clear refutations emerged. The first contribution (latent action module for particle dynamics) examined six candidates with none providing overlapping prior work. The second contribution (state-of-the-art video prediction) examined ten candidates, again with no refutations. The third contribution (goal-conditioned imitation learning application) similarly found no refuting work among ten candidates. This suggests that within the limited search scope, the combination of particle-based representations, stochastic dynamics modeling, and decision-making integration appears relatively unexplored, though the modest candidate pool (26 papers) means substantial relevant work may exist beyond this analysis.

Based on the limited literature search, the work appears to occupy a distinctive position combining particle-based scene decomposition with stochastic world modeling and control applications. The sparse population of its taxonomy leaf and absence of refuting candidates among 26 examined papers suggest novelty, though this assessment is constrained by the search scope. The cross-cutting nature—

spanning representation learning, dynamics prediction, and decision-making—may contribute to the lack of direct precedents, as most prior work focuses on narrower aspects of this pipeline.

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

Contribution 1: Self-supervised object-centric world model with novel latent action module

Description: The authors introduce LPWM, which combines object-centric particle representations with a novel context module that predicts per-particle latent action distributions. This enables stochastic dynamics modeling and supports flexible conditioning on actions, language, image goals, and multi-view inputs, all trained end-to-end from video data.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Learning to Act Anywhere with Task-centric Latent Actions

URL: View paper

Brief Assessment

Task-centric Latent Actions[51] focuses on learning task-centric latent actions for robotic manipulation without explicit object-centric particle representations or stochastic dynamics modeling via per-particle latent action distributions. The architectural approaches and representation learning mechanisms differ fundamentally.

---

2. Object-level Scene Deocclusion

URL: View paper

Brief Assessment

Object-level Scene Deocclusion[54] focuses on scene deocclusion and 3D reconstruction from occluded objects, not on world models for dynamics prediction or latent action learning for sequential decision-making.

---

3. MultiPLY: A Multisensory Object-Centric Embodied Large Language Model in 3D World

URL: View paper

Brief Assessment

MultiPLY[52] focuses on multisensory embodied LLMs for 3D environments with action tokens for agent interaction, not on self-supervised object-centric world models with latent action distributions for stochastic dynamics modeling from video data.

---

4. Narrate2Nav: Real-Time Visual Navigation with Implicit Language Reasoning in Human-Centric Environments

URL: View paper

Brief Assessment

Narrate2Nav[55] focuses on real-time vision-action models for mobile robot navigation using implicit language reasoning in visual encoders, not on self-supervised object-centric world models with latent action distributions for stochastic dynamics modeling as proposed in the original paper.

---

5. Latent Action Pretraining Through World Modeling

URL: View paper

Brief Assessment

Latent Action Pretraining[53] focuses on learning latent action representations from unlabeled video data for imitation learning across tasks and embodiments, but does not propose an object-centric world model with particle representations or per-particle latent action distributions as in the original paper.

---

6. Object-Centric World Model for Language-Guided Manipulation

URL: View paper

Brief Assessment

Object-Centric World Model[14] uses slot attention for object-centric representation and focuses on language-guided manipulation tasks, whereas the original paper uses particle-based representations with per-particle latent actions for general video prediction and decision-making across diverse datasets.

Contribution 2: State-of-the-art object-centric video prediction on diverse datasets

Description: LPWM achieves superior performance compared to existing object-centric methods across multiple real-world robotics datasets and simulated environments, demonstrating improved visual quality metrics and the ability to model complex multi-object interactions while maintaining object permanence.

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. Learning what and where: Disentangling location and identity tracking without supervision

URL: View paper

Brief Assessment

Learning What Where[62] focuses on tracking and localization in synthetic datasets (CATER, Moving MNIST) with explicit object permanence mechanisms, while the original paper addresses stochastic video prediction across real-world robotics datasets with latent action modeling. The technical approaches and evaluation domains differ substantially.

---

2. Hopper: Multi-hop Transformer for Spatiotemporal Reasoning

URL: View paper

Brief Assessment

Hopper[64] focuses on spatiotemporal reasoning and object localization queries in videos, not video prediction or generation. The candidate addresses object permanence through query-based localization rather than predicting future video frames.

---

3. Out of sight, still in mind: Reasoning and planning about unobserved objects with video tracking enabled memory models

URL: View paper

Brief Assessment

Out of Sight[59] focuses on memory-based reasoning for manipulation planning with occluded objects in robotics, not on general object-centric video prediction benchmarks. The candidate addresses object permanence through explicit memory management for planning tasks, while the original contribution concerns video prediction quality metrics across diverse datasets.

---

4. Looping loci: Developing object permanence from videos

URL: View paper

Brief Assessment

Looping Loci[61] focuses on learning object permanence through occlusions in controlled synthetic datasets (ADEPT, CLEVRER), not on achieving state-of-the-art video prediction performance across diverse real-world robotics datasets with complex multi-object interactions as claimed by the original paper.

---

5. Unsupervised learning of object structure and dynamics from videos

URL: View paper

Brief Assessment

Unsupervised Object Structure Dynamics[60] focuses on keypoint-based video prediction with unsupervised learning, while LPWM addresses stochastic dynamics modeling with latent actions and multi-modal conditioning. The candidate does not demonstrate prior work on LPWM's specific contributions regarding latent particle representations, per-particle latent actions, or comprehensive conditioning modalities.

---

6. SlotPi: Physics-informed Object-centric Reasoning Models

URL: View paper

Brief Assessment

SlotPi[56] focuses on physics-informed slot-based models for physical reasoning tasks including fluids and VQA, not on general object-centric video prediction benchmarks with object permanence metrics that LPWM addresses.

---

7. Occlusion resistant learning of intuitive physics from videos

URL: View paper

Brief Assessment

Occlusion Resistant Learning[65] focuses on learning intuitive physics from videos with occlusions using a probabilistic formulation and compositional renderer, not on general object-centric video prediction across diverse datasets with object permanence as the primary contribution.

---

8. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties

URL: View paper

Brief Assessment

Physion++[63] focuses on evaluating physical scene understanding through property inference tasks (mass, friction, elasticity, deformability) rather than advancing object-centric video prediction methods. The candidate is a benchmark/evaluation paper, not a video prediction model.

---

9. Learning object permanence from videos via latent imaginations

URL: View paper

Brief Assessment

Object Permanence Videos[58] focuses on learning object permanence through occlusions in synthetic datasets, not on general multi-object video prediction performance across diverse real-world robotics datasets as claimed by the original paper.

---

10. Physion: Evaluating physical prediction from vision in humans and machines

URL: View paper

Brief Assessment

Physion[57] focuses on evaluating physical prediction capabilities through binary contact prediction tasks in physics simulation scenarios, not on video prediction performance metrics or object permanence modeling as claimed in the original contribution.

Contribution 3: Application to decision-making via goal-conditioned imitation learning

Description: The authors show that pre-trained LPWM can be adapted for goal-conditioned imitation learning by learning a simple mapping from latent actions to real actions. They demonstrate competitive performance on multi-object manipulation tasks, establishing LPWM's practical utility beyond video prediction.

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. The Art of Imitation: Learning Long-Horizon Manipulation Tasks From Few Demonstrations

URL: View paper

Brief Assessment

Art of Imitation[73] focuses on task-parametrized Gaussian mixture models for manipulation from demonstrations, not on adapting pre-trained world models with latent actions for goal-conditioned imitation learning as in the original paper.

---

2. ForceMimic: Force-Centric Imitation Learning with Force-Motion Capture System for Contact-Rich Manipulation

URL: View paper

Brief Assessment

ForceMimic[75] focuses on force-centric imitation learning for contact-rich manipulation tasks like vegetable peeling, not on goal-conditioned imitation learning using world models for multi-object manipulation as described in the original contribution.

---

3. LUMOS: Language-Conditioned Imitation Learning with World Models

URL: View paper

Brief Assessment

LUMOS[71] focuses on language-conditioned imitation learning with world models for robotics, while the original paper demonstrates goal-conditioned imitation learning using image-based goals with pre-trained LPWM. The technical approaches differ significantly in conditioning modalities and training procedures.

---

4. Goal-Conditioned Imitation Learning using Score-based Diffusion Policies

URL: View paper

Brief Assessment

Goal-Conditioned Diffusion Policies[66] focuses on goal-conditioned imitation learning using score-based diffusion models for policy representation, not on adapting pre-trained world models with latent actions for manipulation tasks as in the original paper.

---

5. ICRT: In-Context Imitation Learning via Next-Token Prediction

URL: View paper

Brief Assessment

ICRT[69] focuses on in-context imitation learning through next-token prediction on sensorimotor trajectories, not on adapting pre-trained world models for goal-conditioned tasks via latent action mappings as in the original paper.

---

6. In-context imitation learning via next-token prediction

URL: View paper

Brief Assessment

In-context Next-Token[67] focuses on in-context learning for robotics using sensorimotor trajectory prompts, not on adapting pre-trained world models for goal-conditioned imitation learning via latent action mappings as in the original paper.

---

7. One-shot imitation learning with graph neural networks for pick-and-place manipulation tasks

URL: View paper

Brief Assessment

One-shot Graph Networks[70] focuses on pick-and-place manipulation tasks using GNN-based policies with synthetic demonstrations, not on adapting pre-trained world models for goal-conditioned imitation learning in multi-object manipulation environments.

---

8. Visual Imitation Learning of Task-Oriented Object Grasping and Rearrangement

URL: View paper

Brief Assessment

Visual Imitation Grasping[74] focuses on task-oriented object grasping and rearrangement using implicit neural fields for manipulation tasks, not on general world models with latent actions for multi-object manipulation as in the original paper.

---

9. Data Scaling Laws in Imitation Learning for Robotic Manipulation

URL: View paper

Brief Assessment

Data Scaling Laws[68] focuses on data scaling principles for imitation learning in robotic manipulation, not on world model architectures or latent action learning for goal-conditioned tasks. The candidate does not address world models or latent particle representations.

---

10. SCIL: Stage-Conditioned Imitation Learning for Multi-Stage Manipulation

URL: View paper

Brief Assessment

SCIL[72] focuses on stage-conditioned imitation learning for multi-stage manipulation tasks with ambiguity resolution between stages, not on goal-conditioned imitation learning using world models for multi-object manipulation as in the original paper.

Appendix: Text Similarity Detection

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

References

- [0] Latent Particle World Models: Self-supervised Object-centric Stochastic Dynamics Modeling View paper
- [1] Object-centric Video Prediction without Annotation View paper
- [2] Robot Learning from a Physical World Model View paper
- [3] Spatial policy: Guiding visuomotor robotic manipulation with spatial-aware modeling and reasoning View paper
- [4] DriVerse: Navigation World Model for Driving Simulation via Multimodal Trajectory Prompting and Motion Alignment View paper
- [5] Physics-Grounded Motion Forecasting via Equation Discovery for Trajectory-Guided Image-to-Video Generation View paper
- [6] SlotDiffusion: Object-Centric Generative Modeling with Diffusion Models View paper
- [7] PlaySlot: Learning Inverse Latent Dynamics for Controllable Object-Centric Video Prediction and Planning View paper
- [8] UniUGP: Unifying Understanding, Generation, and Planing For End-to-end Autonomous Driving View paper
- [9] Scalor: Generative world models with scalable object representations View paper
- [10] Physgen: Rigid-body physics-grounded image-to-video generation View paper
- [11] SCOPE: Stochastic Cartographic Occupancy Prediction Engine for Uncertainty-Aware Dynamic Navigation View paper
- [12] GraphMimic: Graph-to-Graphs Generative Modeling from Videos for Policy Learning View paper
- [13] Time-Conditioned Generative Modeling of Object-Centric Representations for Video Decomposition and Prediction View paper
- [14] Object-Centric World Model for Language-Guided Manipulation View paper
- [15] GEM: A Generalizable Ego-Vision Multimodal World Model for Fine-Grained Ego-Motion, Object Dynamics, and Scene Composition Control View paper
- [16] Deep variational Luenberger-type observer with dynamic objects channel-attention for stochastic video prediction View paper
- [17] Agentic World Modeling for 6G: Near-Real-Time Generative State-Space Reasoning View paper
- [18] Wcdt: World-Centric Diffusion Transformer for Traffic Scene Generation View paper
- [19] Scalable Multi-Temporal Remote Sensing Change Data Generation via Simulating Stochastic Change Process View paper
- [20] Stochastic prediction of multi-agent interactions from partial observations View paper
- [21] Slotformer: Unsupervised visual dynamics simulation with object-centric models View paper
- [22] Advances in object-centric learning methods towards real world applications View paper
- [23] Generative video transformer: Can objects be the words? View paper

- [24] TextOCVP: Object-Centric Video Prediction with Language Guidance View paper
- [25] Learning Object-Centric Representations Based on Slots in Real World Scenarios View paper
- [26] Self-supervised multi-future occupancy forecasting for autonomous driving View paper
- [27] PLM-based World Models for Text-based Games View paper
- [28] Object-Centric Dreamer View paper
- [29] Predicting human activities using stochastic grammar View paper
- [30] Learning temporal transformations from time-lapse videos View paper
- [31] Interaction Aware Relational Representations for Video Prediction View paper
- [32] OCK: Unsupervised Dynamic Video Prediction with Object-Centric Kinematics View paper
- [33] GenTrack: A New Generation of Multi-Object Tracking View paper
- [34] Compositional video prediction View paper
- [35] SOLD: Slot Object-Centric Latent Dynamics Models for Relational Manipulation Learning from Pixels View paper
- [36] Seeing the Future, Perceiving the Future: A Unified Driving World Model for Future Generation and Perception View paper
- [37] VFMF: World Modeling by Forecasting Vision Foundation Model Features View paper
- [38] DyMoDreamer: World Modeling with Dynamic Modulation View paper
- [39] A Symmetric and Object-Centric World Model for Stochastic Environments View paper
- [40] SAMPO:Scale-wise Autoregression with Motion PrOmpt for generative world models View paper
- [41] 3D Video Models through Point Tracking, Reconstructing and Forecasting View paper
- [42] AXIOM: Learning to Play Games in Minutes with Expanding Object-Centric Models View paper
- [43] WorldGym: World Model as An Environment for Policy Evaluation View paper
- [44] Effectively Indexing Uncertain Moving Objects for Predictive Queries View paper
- [45] Natural Building Blocks for Structured World Models: Theory, Evidence, and Scaling View paper
- [46] Generative World Modelling for Humanoids: 1X World Model Challenge Technical Report View paper
- [47] WorldDreamer: Towards General World Models for Video Generation via Predicting Masked Tokens View paper
- [48] Patch-based Object-centric Transformers for Efficient Video Generation View paper
- [49] Pandora: Towards General World Model with Natural Language Actions and Video States View paper
- [50] Learning Physical Dynamics for Object-centric Visual Prediction View paper
- [51] Learning to Act Anywhere with Task-centric Latent Actions View paper
- [52] MultiPLY: A Multisensory Object-Centric Embodied Large Language Model in 3D World View paper
- [53] Latent Action Pretraining Through World Modeling View paper
- [54] Object-level Scene Deocclusion View paper
- [55] Narrate2Nav: Real-Time Visual Navigation with Implicit Language Reasoning in Human-Centric Environments View paper
- [56] SlotPi: Physics-informed Object-centric Reasoning Models View paper
- [57] Physion: Evaluating physical prediction from vision in humans and machines View paper
- [58] Learning object permanence from videos via latent imaginations View paper
- [59] Out of sight, still in mind: Reasoning and planning about unobserved objects with video tracking enabled memory models View paper
- [60] Unsupervised learning of object structure and dynamics from videos View paper
- [61] Looping loci: Developing object permanence from videos View paper
- [62] Learning what and where: Disentangling location and identity tracking without supervision View paper
- [63] Physion++: Evaluating physical scene understanding that requires online inference of different physical properties View paper
- [64] Hopper: Multi-hop Transformer for Spatiotemporal Reasoning View paper
- [65] Occlusion resistant learning of intuitive physics from videos View paper
- [66] Goal-Conditioned Imitation Learning using Score-based Diffusion Policies View paper
- [67] In-context imitation learning via next-token prediction View paper
- [68] Data Scaling Laws in Imitation Learning for Robotic Manipulation View paper
- [69] ICRT: In-Context Imitation Learning via Next-Token Prediction View paper
- [70] One-shot imitation learning with graph neural networks for pick-and-place manipulation tasks View paper
- [71] LUMOS: Language-Conditioned Imitation Learning with World Models View paper
- [72] SCIL: Stage-Conditioned Imitation Learning for Multi-Stage Manipulation View paper
- [73] The Art of Imitation: Learning Long-Horizon Manipulation Tasks From Few Demonstrations View paper
- [74] Visual Imitation Learning of Task-Oriented Object Grasping and Rearrangement View paper
- [75] ForceMimic: Force-Centric Imitation Learning with Force-Motion Capture System for Contact-Rich Manipulation View paper