OCK: Unsupervised Dynamic Video Prediction with Object-Centric Kinematics

Yeon-Ji Song[1,2], Jaein Kim[1,2*], Suhyung Choi[1,2*], Jin-Hwa Kim[2,3†], Byoung-Tak Zhang[1,2†]
Seoul National University[1] SNU AIIS[2] NAVER AI Lab[3]

ICCV OCT 19-23, 2025 HONOLULU HAWAII

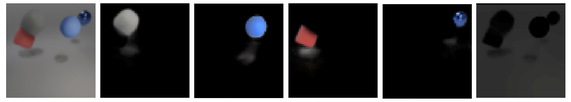
TL;DR

OCK is an object-centric video prediction model that leverages Object Kinematics to understand time-varying motion alongside time-static object appearance. By incorporating explicit kinematic attributes, OCK effectively models spatiotemporal object patterns, leading to superior generalization and more accurate long-term predictions in dynamic and complex scenarios.

Introduction

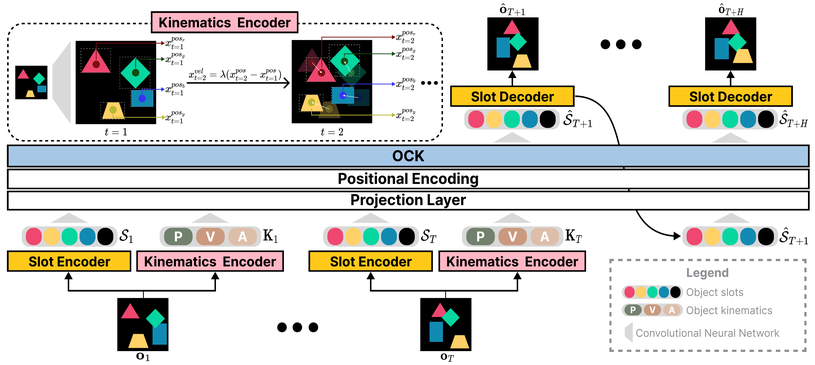
- We introduce **Object-Centric Kinematics (OCK)**, a novel method that explicitly integrates object appearances and object kinematics.
- OCK's unique approach accurately models spatiotemporal interactions by leveraging object-centric motion attributes.
- By overcoming the limitations of models that rely solely on static visual attributes, our integrated approach achieves superior performance, particularly in complex and dynamic environments.

Motivation



- Object-centric representation is a paradigm that decomposes a scene into discrete, localized *slots*, enabling structured reasoning, compositionality, and generalization based on individual object properties and relationships.
- Human perception understands scenes by decomposing them into **time-static object appearances** and **time-varying object motions**.
- This object-centric approach offers a powerful inductive bias for such video prediction, enhancing accuracy and enabling precise modeling of complex physical interactions (e.g., collisions, freefall) that challenge general transformer-based models.
- However, existing object-centric models often overlook explicit motion dynamics in favor of preserving appearances, which limits their ability to capture complex interactions and maintain temporal consistency.

Method



OCK is built upon an autoregressive object-centric transformer, which uses two parallel encoding modules to predict future frames:

- **Slot Encoder**: We use SAVi[1] model to extract permutation-invariant **object slots** $\mathcal{S}_t$ from video frames, which capture object appearances.
- **Kinematics Encoder**: A CNN-based network extracts explicit object motion information named **Object Kinematics** $\mathbf{K}_t$, including position, velocity, and acceleration:

$$\mathbf{K}_t \triangleq \begin{bmatrix} \mathrm{x}_t^{\mathrm{pos}} \\ \mathrm{x}_t^{\mathrm{vel}} \\ \mathrm{x}_t^{\mathrm{acc}} \end{bmatrix} = \begin{bmatrix} \phi(o_t) \\ \lambda(\mathrm{x}_t^{\mathrm{pos}} - \mathrm{x}_{t-1}^{\mathrm{pos}}) \\ \mathrm{x}_t^{\mathrm{vel}} - \mathrm{x}_{t-1}^{\mathrm{vel}} \end{bmatrix}$$

These features are subsequently fed into the OCK transformer. We introduce two alternative architectures:

- **Joint-OCK:** This model concatenates object slots and kinematics as a single input for a standard transformer.
- **Cross-OCK**: This model leverages the cross-attention mechanism with object slots act as queries and Object Kinematics as keys and values to enhance computational efficiency.

The model is trained in two steps: first by pretraining a SAVi model to decompose video frames into object slots, and second, by training the OCK model to predict future slots using a combined loss for both object and image reconstruction.

1 SAVi stands for Slot Attention for Video. See Kipf et al. 2021 for details.

Results

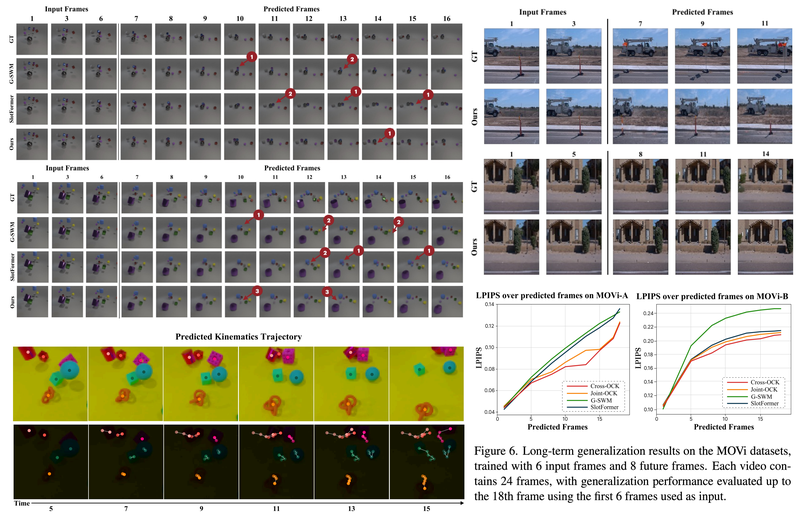
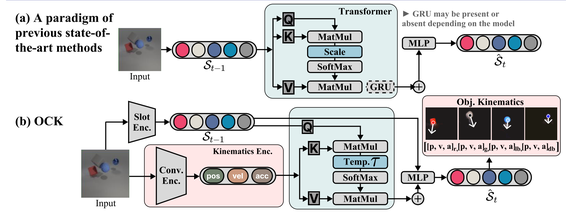


Figure 6. Long-term generalization results on the MOVi datasets, trained with 6 input frames and 8 future frames. Each video contains 24 frames, with generalization performance evaluated up to the 18th frame using the first 6 frames used as input.

Model	OBJ3D PSNR↑	OBJ3D SSIM↑	OBJ3D LPIPS↓	MOVi-A PSNR↑	MOVi-A SSIM↑	MOVi-A LPIPS↓	MOVi-B PSNR↑	MOVi-B SSIM↑	MOVi-B LPIPS↓	MOVi-C PSNR↑	MOVi-C SSIM↑	MOVi-C LPIPS↓	MOVi-D PSNR↑	MOVi-D SSIM↑	MOVi-D LPIPS↓	MOVi-E PSNR↑	MOVi-E SSIM↑	MOVi-E LPIPS↓
G-SWM	31.142	0.900	0.039	26.140	0.784	0.133	21.850	0.677	0.247	19.466	0.451	0.554	20.567	0.548	0.355	21.166	0.534	0.359
SlotFormer	33.083	0.932	0.024	25.180	0.785	0.134	21.329	0.690	0.215	19.482	0.456	0.534	20.675	0.565	0.332	21.269	0.547	0.335
OCVP-Seq	33.100	0.932	0.025	26.240	0.789	0.127	**21.978**	0.701	0.219	17.945	0.415	0.631		Diverge			Diverge	
OCVP-Par	32.990	0.931	0.025	26.310	0.788	0.127	21.909	0.688	0.226	17.941	0.402	0.650		Diverge			Diverge	
Joint-OCK	**35.125**	**0.958**	**0.019**	27.259	0.811	0.124	21.646	0.695	**0.198**	21.038	**0.593**	**0.370**	22.087	0.557	0.282	22.394	0.569	**0.302**
Cross-OCK	34.097	0.925	**0.019**	**27.576**	**0.812**	**0.123**	21.482	**0.703**	0.209	**21.040**	0.592	0.376	**22.338**	**0.568**	**0.236**	**22.340**	**0.572**	**0.302**

Table 1. Evaluation of video prediction quality across six synthetic datasets, increasing in scene complexity from left to right. "Diverge" denotes the phenomenon where prediction performance degrades due to suboptimal slot extraction by the encoder.

Previous work vs. Ours



Code

github.com/yeonjisong/OCK

Any queries?
Contact me at
yjsong@snu.ac.kr