

OCK: Unsupervised Dynamic Video Prediction with Object-Centric Kinematics

Yeon-Ji Song^{1,2}, Jaemin Kim^{1,2*}, Suhyung Choi^{1,2*}, Jin-Hwa Kim^{2,3†}, Byoung-Tak Zhang^{1,2†}

¹Seoul National University ²SNU AIIS ³NAVER AI Lab

{yjsong, jykim, s.choi}@snu.ac.kr, jlnhwa.kim@navercorp.com, btzhang@snu.ac.kr

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, they primarily focus on object appearance, often overlooking motion dynamics, which is crucial for modeling dynamic interactions and maintaining temporal consistency in complex environments. To address these limitations, we propose OCK, a dynamic video prediction model leveraging object-centric kinematics and object slots. We introduce a novel component named Object Kinematics that comprises explicit object motions, serving as an additional attribute beyond conventional appearance features to model dynamic scenes. The Object Kinematics are integrated into various OCK mechanisms, enabling spatiotemporal prediction of complex object interactions over long video sequences. Our model demonstrates superior performance in handling complex scenes with intricate object attributes and motions, highlighting its potential for applicability in vision-related dynamics learning tasks.

1. Introduction

Human-level intelligence requires a deep understanding of the environment and the spatiotemporal interactions of surrounding objects [13]. This proficiency is fundamental to deep learning, as it underpins the human ability to perceive scenes as structured compositions of object components, facilitating object recognition in dynamically changing environments [17]. Despite advancements in modeling human intelligence, current approaches struggle to bridge the gap between human-level understanding and machine-

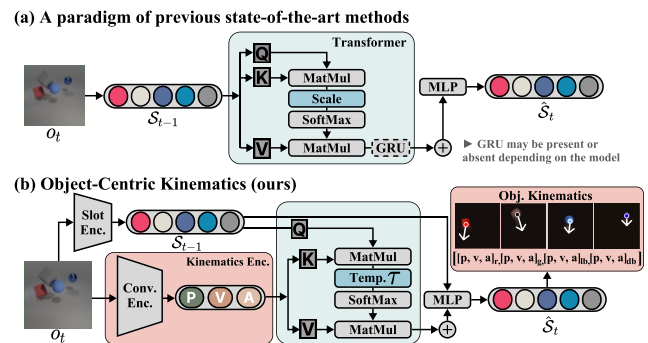


Figure 1. Comparison of OCK with previous slot-based video prediction methods utilizing transformers. The previous methods leverage object attributes only. In contrast, OCK utilizes both object appearances and motions.

generated world models, lacking the ability to accurately replicate human perception and interpretation of their surroundings. To address this limitation, unsupervised object-centric representation learning has emerged as a promising framework for capturing the compositional structure of objects, mirroring the human ability to intuitively interpret their surroundings. This approach has shown particular effectiveness in dynamic environments with diverse, constantly evolving, or previously unseen objects, achieving remarkable performance across various domains, including scene understanding [31, 35], object tracking [16, 48], reinforcement learning [45, 46] and video prediction [21, 44].

Our work focuses on object-centric video prediction. The object-centric bias significantly improves the model’s ability to capture object dynamics and interactions for successful video prediction, leading to enhanced predictive accuracy in complex environments. This approach optimally balances between expressivity and accuracy [7], enabling precise modeling of complex physical interactions (i.e., collision, freefall, mass and charge) that remain challenging for more general models, despite imposing constraints on the range of visual complexity it can effectively represent.

Among various approaches to object-centric video prediction, transformers have demonstrated efficiency in ad-

*Equal contribution

†Corresponding authors

addressing the trade-off between expressivity and long-term predictive accuracy through attention mechanisms and sequential positional encoding [36]. This functionality of transformers combined with object-centric representation enables direct modeling of the spatiotemporal dynamics of objects within the object representation space [37, 44]. While prior works have achieved exceptional performance, certain approaches are restricted to simplistic environments, focusing solely on implicit object appearances [21], while others can handle complex scenes but are often limited to short videos [37, 44]. See Fig. 1 for a visual explanation. Successful prediction fundamentally relies on capturing *long-term dependencies* and *scene complexity*, including object interactions, dynamics, and visual variations. The absence of either factor in prior works highlights the need for a model that effectively captures explicit physical interactions and temporal dependencies, with an emphasis on precise object motion modeling within video frames.

In this paper, we present Object-Centric Kinematics (OCK) for unsupervised object-centric video prediction. We introduce Object Kinematics, which encapsulates object motion within a structured image space and plays a pivotal role in object tracking, demonstrating efficacy across both synthetic and real-world environments. The Object Kinematics are derived through two approaches: The *analytical approach* anticipates subsequent states from the input frame via logical reasoning and utilizes them for subsequent frame generation; The *empirical approach* leverages the given object information via inductive reasoning, focusing on its implicit learning. We employ two versions of OCK to integrate the Object Kinematics with object appearances, named slots [22]. We compare our model with the prior works that utilize object appearance information only [21, 44] in complex environments where object motion, appearance, and backgrounds vary repeatedly.

In summary, we make the following contributions: (1) We introduce an object-centric video prediction model leveraging Object Kinematics to comprehend time-varying object motion and time-static object appearance. (2) We conduct comprehensive experiments with four recent slot-based methods on six synthetic and one real-world datasets. Our model significantly enhances the accuracy of predicting object dynamics in highly complex scenarios and exhibits better generalization to long sequences during test time. (3) We investigate the impact and necessity of Object Kinematics through various ablation studies.

2. Related work

2.1. Object-centric representation learning

Recent unsupervised object-centric learning can be split into three approaches. Spatial attention approaches employ CNN or spatial transformer networks to crop rectangular

regions from an image, enabling the extraction of object attributes such as position, scale, or latent [2, 4, 5, 11, 19]. However, these approaches often rely on a fixed-size sampling grid or coarse bounding box, which may not be suitable for scenes featuring diverse object sizes, potentially compromising training efficacy when the sampling grid fails to overlap with any object. Sequential attention models, exemplified by RNN-based frameworks, sequentially attend to different regions in an image [3, 9], resulting in a suboptimal understanding of interrelationships between objects and image regions. As a result, these models struggle to capture the global context of the image. Lastly, iterative attention methods initialize a set of object representations, namely *object slots*, and iteratively refine them to associate them with distinct regions of an image [30, 31, 40]. Predominantly inspired by Slot Attention [22], the iterative approach fosters competition among object slots by employing attention along the object dimension. Our work focuses on the iterative attention approach, given its widespread usage for video prediction tasks.

2.2. Dynamic video prediction

Video prediction and generation in dynamic environments is a challenging task that has gained significant attention in recent years. Approaches in video modeling include object-agnostic models, utilizing 3D convolutions [10, 38] or RNNs [25, 39], and those employing structured or object-centric models, using probabilistic modeling techniques [35], transformers [32, 37, 44], or 3D point clouds [33]. The object-agnostic models often require explicit human supervision, may be restricted to simple 2D datasets, and typically concentrate on modeling temporal changes using image features. These approaches neglect explicit consideration of the composition of video frames. In contrast, approaches utilizing object-centric representations to disentangle frames into object attributes provide a more nuanced and comprehensive understanding beyond general representations. Despite advancements in video prediction, generalization remains limited, with significant challenges in applying these models to unannotated data [43]. Inspired by video prediction models utilizing object-centric representations, our work diverges by investigating transformers' input components, architecture, and functioning.

2.3. Object-centric video prediction with transformers

Our research builds on the foundations of SlotFormer [44] and OCVP [37], both rooted in transformer architectures, to advance object-centric video prediction and generation. OCVP extends SlotFormer by slightly modifying the spatiotemporal attention block in two different ways. The recurrent transformer strategy they utilize facilitates long-term predictions while mitigating the information loss on

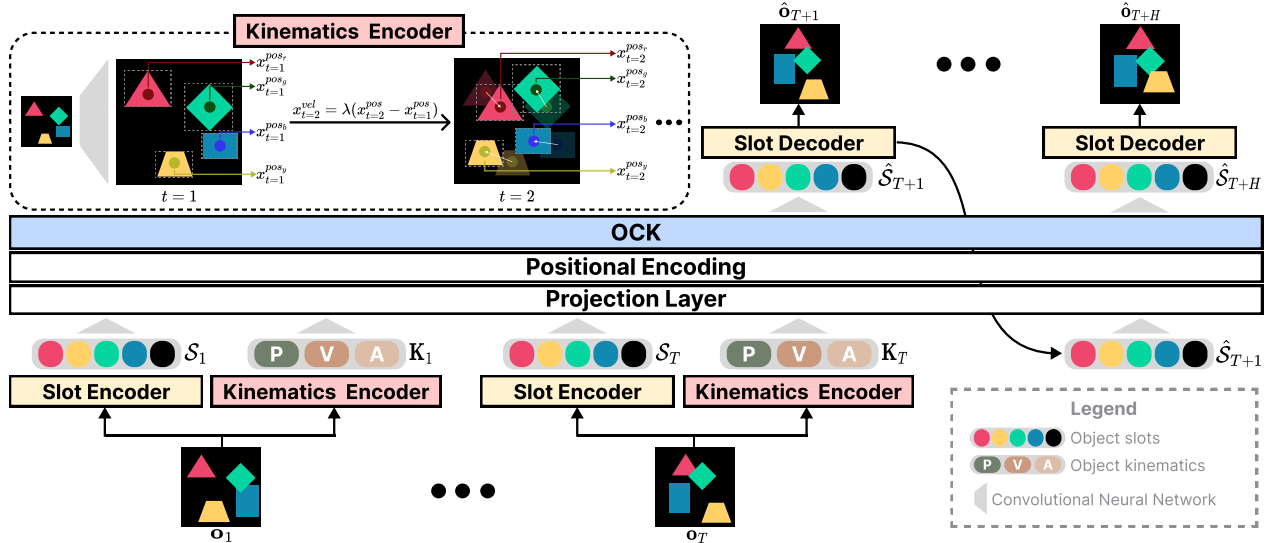


Figure 2. **Overview of our pipeline.** The model extracts object slots using the pretrained object-centric slot encoder, and Object Kinematics, including position, velocity, and acceleration, using the kinematics encoder, both from multiple video frames. These features are linearly projected and processed by the OCK mechanism with a positional encoding to generate future slots in an autoregressive manner.

spatial object motion [29]. Additionally, their use of object-centric representations introduces an inductive bias for understanding object attributes within a sequence [23]. However, their object-wise representations are biased toward invariant visual cues, focusing narrowly on implicit changes in video frames. This limited scope results in a lack of depth in capturing advanced object motions. Extensive research has explored the generalization capabilities of transformers, particularly in symbolic mathematical modeling and compositional reasoning [8, 24, 42]. Building upon this area, our paper proposes a novel framework that comprehends spatiotemporal object dynamics and predicts the explicit sequence of time-varying object motions and time-static object appearances across various time steps.

3. Method

OCK is built upon an autoregressive object-centric transformer as summarized in Fig. 2. The input to the OCK is obtained through two modules, involving the extraction of object slots and Object Kinematics. We demonstrate how the two features are subsequently fed into the OCK transformers, forming the basis for spatiotemporal future frame prediction and generation in an autoregressive way.

3.1. Slot encoder

We leverage the Slot Attention [22] framework to map image features to permutation-invariant object slots. A convolutional neural network extracts a grid of image features from a video frame \mathbf{o}_t , which are then flattened into a set of feature map $\mathbf{h}_t \in \mathbb{R}^{M \times D_{\text{fl}}}$ that represents its semantic information, where M is the size of the flattened $H \times W$ feature

grid and D_{fl} is the dimensionality of the extracted feature maps. At each iteration, the model randomly initializes N object slots $\mathcal{S}_0 \in \mathbb{R}^{N \times D_{\text{slot}}}$ and performs Slot Attention to update the object slots via iterative Scaled Dot-Product Attention [36]. This attention mechanism computes softmax over object slots and updates the slots with the weighted average to represent a part of the input. The output \mathcal{S}_t is a set of object slots, which is fed into the autoregressive transformer along with the Object Kinematics.

3.2. Kinematics encoder

We simultaneously perform Object Kinematics extraction, which involves a convolutional neural network (CNN) and numerical calculations to estimate the object motions. The intuition behind this is to preserve the explicit properties of objects, thereby preventing the model from learning incorrect motions in the initial stages of training when scaling up to complex environments, which is a phenomenon often observed when relying solely on slots [21, 44].

Initially, video frame \mathbf{o}_t is processed by a CNN-based slot network ϕ to extract a grid of low-level image features, which are then utilized to generate low-dimensional object features \mathbf{x}_t by localizing the center of mass of each object as 2D coordinates. The object features are passed through an MLP to ensure consistent encodings, forming the *position* state denoted as $\mathbf{x}_t^{\text{pos}}$. Note that $\mathbf{x}_t^{\text{pos}}$ is equivariant to the permutation over objects in \mathbf{o}_t . The *velocity* state $\mathbf{x}_t^{\text{vel}}$ is derived by computing the difference between two consecutive position states, $\mathbf{x}_t^{\text{pos}}$ and $\mathbf{x}_{t-1}^{\text{pos}}$. The *acceleration* state $\mathbf{x}_t^{\text{acc}}$ is calculated via the differences between $\mathbf{x}_t^{\text{vel}}$ and $\mathbf{x}_{t-1}^{\text{vel}}$. To ensure the consistency in the unified kinematic representation,

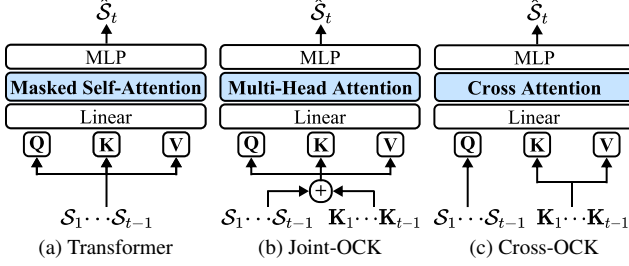


Figure 3. An illustration of the OCK architecture. (a) Traditional transformer model processes slot exclusively. (b) Joint-OCK incorporates concatenated slots and kinematics as input. (c) Cross-OCK employs cross-attention, where slot representations serve as queries, and kinematics functions as keys and values.

a learnable parameter λ is introduced to scale the velocity relative to the position, formulated as follows:

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

The Object Kinematics $\mathbf{K}_t \in \mathbb{R}^{N \times D_{\text{kin}}}$ is constructed by concatenating the low-level geometric object states and is subsequently fed as input to the transformer module along with the object slots \mathcal{S}_t . Notably, Object Kinematics is learned through the objects’ centers extracted from the slot encoder, without reliance on task-specific loss functions, such as frame reconstruction or state transition objectives. Modeling kinematics in 2D image space is deliberate, as extending it to a 3D depth image requires an additional network, leading to high computational costs due to the reliance on a pretrained optical flow or frame reconstruction model [14, 18]. Furthermore, the predictive performance remains robust even in the absence of rotation and scaling factors. Accordingly, the Object Kinematics are employed in two ways: *analytical* and *empirical* approach.

Analytical approach leverages Object Kinematics from the video frame at time t to predict the subsequent position state $\mathbf{x}_{t+1}^{\text{pos}'}$ based on the premise that objects will continue their motion patterns shortly. OCK integrates current kinematics $\mathbf{x}_t^{\text{pos}}$ with the anticipated kinematics $\mathbf{x}_{t+1}^{\text{pos}'}$ to generate the subsequent frame $\hat{\mathbf{x}}_{t+1}^{\text{pos}}$. This approach leverages predicted kinematic information to emphasize continuous motion patterns through a reasoning-driven method.

Empirical approach relies solely on the current video frame \mathbf{x}_t processed through an inductive gradient-based learning without additional dependencies. Specifically, it extracts Object Kinematics \mathbf{K}_t from the video frame similarly, but focusing exclusively on the kinematics information at time t . The motion pattern inferred from the current frame is utilized to predict the subsequent frame $\hat{\mathbf{x}}_{t+1}^{\text{pos}}$.

Remark: The analytical approach is grounded in structured

Algorithm 1 Object-Centric Kinematics (OCK)

```

1: Inputs: input frames  $\text{input}$ , learnable queries  $\text{init}$ ,
   number of iterations  $T$ 
2: Require: time difference  $\delta$ , learnable parameter  $\lambda$ 
3: Modules: OCK module  $\text{OCK}(\cdot, \cdot)$ , object slot module
    $\text{SA}(\cdot, \cdot)$ , Object Kinematics module  $\text{KIN}(\cdot)$ 
4:  $\text{slots} \leftarrow \text{init}$ 
5: for  $i \leftarrow 1$  to  $T$  do
6:    $\text{slots} \leftarrow \text{SA}(\text{slots}, \text{input})$ 
7:    $\text{kins} \leftarrow \text{KIN}(\text{input})$ 1
8:   if Analytical Approach then
9:      $\text{kins}'_{\text{pos}} \leftarrow \text{kins}_{\text{pos}} + \text{kins}_{\text{vel}} \times \delta$ 
10:     $\text{kins}'_{\text{vel}} \leftarrow \lambda(\text{kins}'_{\text{pos}} - \text{kins}_{\text{pos}})$ 
11:     $\text{kins}'_{\text{acc}} \leftarrow \text{kins}'_{\text{vel}} - \text{kins}_{\text{vel}}$ 
12:     $\text{kins}' \leftarrow [\text{kins}'_{\text{pos}}, \text{kins}'_{\text{vel}}, \text{kins}'_{\text{acc}}]$ 
13:     $\text{slots}' \leftarrow \text{OCK}(\text{slots}, [\text{kins}, \text{kins}'])$ 
14:   else if Empirical Approach then
15:      $\text{slots}' \leftarrow \text{OCK}(\text{slots}, \text{kins})$ 
16:   end if
17:    $\text{slots} \leftarrow \text{slots}'$ 
18: end for
19: return  $\text{slots}'$ 

```

logical reasoning, while the empirical approach employs a gradient-based learning as illustrated in Algorithm 1.

3.3. Autoregressive OCK transformers

OCK leverages the object slots \mathcal{S}_t and the Object Kinematics \mathbf{K}_t to predict object slots at the subsequent timestep $\hat{\mathcal{S}}_{t+1}$, while preserving temporal consistency. In this work, we introduce two variants of the OCK mechanism, *Joint-OCK* and *Cross-OCK* as shown in Fig. 3.

Joint-OCK. Similar to Wu et al. [44], Joint-OCK adopts a standard transformer encoder. As shown in Fig. 3b, the input formed by concatenating object slots and Object Kinematics undergoes a linear projection to align with the transformer’s inner dimension d_k , followed by temporal positional encodings to preserve permutation equivariance, ensuring that object slots at the same timestep share the positional encodings for consistent object representations.

$$\mathbf{q} = \mathbf{X}_t \mathbf{W}_q, \quad \mathbf{k} = \mathbf{X}_t \mathbf{W}_k, \quad \mathbf{v} = \mathbf{X}_t \mathbf{W}_v, \quad (2)$$

$$\text{s.t. } \mathbf{X}_t = [\mathcal{S}_t, \mathbf{K}_t],$$

$$\text{Joint-OCK}(\mathbf{v}, \mathbf{k}, \mathbf{q}) = \mathbf{v} \cdot \text{softmax} \left(\frac{\mathbf{k}^\top \mathbf{q}}{\sqrt{d_k}} \right).$$

Cross-OCK. Alternatively, Cross-OCK utilizes a cross-attention mechanism to integrate object slots \mathcal{S}_t with the Object Kinematics \mathbf{K}_t as illustrated in Fig. 3c. Inspired by transformer-based multi-scale feature learning [6], object slots with larger token sizes as queries and Object

¹Note that $\text{kins} = [\text{kins}_{\text{pos}}, \text{kins}_{\text{vel}}, \text{kins}_{\text{acc}}]$ of the current frame.

Kinematics with smaller token sizes as `keys` and `values` achieve enhanced computational efficiency, thereby generating plausible future video frames. The softmax operation is applied along the last dimension to capture the significance of each `key-value` pair for every `query`. The values are then aggregated via a weighted sum to generate the predicted frames. Unlike prior works that scale down weights by `keys`, our framework uses a temperature parameter τ to modulate inner products before softmax, ensuring precise attention calibration [1] as follows:

$$\mathbf{q} = \mathcal{S}_t \mathbf{W}_q, \quad \mathbf{k} = \mathbf{K}_t \mathbf{W}_k, \quad \mathbf{v} = \mathbf{K}_t \mathbf{W}_v, \quad (3)$$

$$\text{Cross-OCK}(\mathbf{v}, \mathbf{k}, \mathbf{q}; \tau) = \mathbf{v} \cdot \text{softmax}\left(\frac{\mathbf{k}^\top \cdot \mathbf{q}}{\tau}\right).$$

The following attention strategies allow the model to handle complex datasets effectively, ensuring robust performance. This proficiency enhances the model’s ability to capture object interactions, discern intricate patterns, and generate comprehensive future frames. Please refer to the supplementary material for a detailed description.

3.4. Model training

OCK is trained in two steps utilizing a pretrained SAVi [20] model. Initially, we train SAVi to decompose video frames into object slots. Then, our model is trained by feeding the extracted object slots into OCK to predict future slots. The predicted slots are transformed into images and masks using the pretrained SAVi for frame reconstruction.

Our model undergoes training by taking the last N output slots from the transformer. These features are then fed into a linear layer to output object slots at the subsequent timestep. To ensure the continuity of future frame prediction, following prior works [37, 44], the generated object slots $\hat{\mathcal{S}}_{T+1}$ serve as input for the subsequent slot prediction $\hat{\mathcal{S}}_{T+2}$ in an autoregressive way, facilitating the generation of any number of future frames H . The model objective is to minimize the object reconstruction loss and image reconstruction loss with a hyperparameter α as follows:

$$\mathcal{L} = \mathcal{L}_{\text{object}} + \alpha \mathcal{L}_{\text{image}}. \quad (4)$$

The object reconstruction loss $\mathcal{L}_{\text{object}}$ is computed as the L2 loss between the ground-truth object slots \mathcal{S}_{T+h} and the reconstructed object slots $\hat{\mathcal{S}}_{T+h}$ as:

$$\mathcal{L}_{\text{object}} = \frac{1}{N \cdot H} \sum_{n=1}^N \sum_{h=1}^H \|\hat{\mathcal{S}}_{T+h}^n - \mathcal{S}_{T+h}^n\|_2. \quad (5)$$

The image reconstruction loss $\mathcal{L}_{\text{image}}$ is calculated using a frozen SAVi decoder f_θ^{SAVi} to transform the predicted object slots $\hat{\mathcal{S}}_{T+h}$ into an image, which is then compared to the corresponding ground-truth video frame \mathbf{o}_{T+h} as:

$$\mathcal{L}_{\text{image}} = \frac{1}{H} \sum_{h=1}^H \|f_\theta^{\text{SAVi}}(\hat{\mathcal{S}}_{T+h}) - \mathbf{o}_{T+h}\|_2. \quad (6)$$

4. Experiments

This section evaluates OCK on synthetic and real-world datasets with diverse scene complexities. We provide an overview of the datasets and baselines, followed by an evaluation of video prediction and scene decomposition accuracy. Then, we conduct ablation studies to investigate the impact of architectural variations on overall performance.

4.1. Experimental details

4.1.1. Datasets

We evaluate the performance of our proposed module across synthetic and real-world datasets to ensure a comprehensive assessment across a spectrum of complex scenarios. **OBJ3D** [21] consists of 3D geometric objects on a gray background, set in motion to simulate dynamic environments in synthetic scenes. Each video consists of 100 frames at a 128×128 resolution, with randomly colored objects rolling toward dynamically positioned central objects. **Multi-Object Video (MOVi)** [12] progressively increases in complexity from A to E, introducing a wider variety of objects’ appearances, motions, and backgrounds. MOVi datasets are specifically designed to test object discovery and tracking in dynamic environments. From an object-centric learning perspective, MOVi- $\{C,D,E\}$ are regarded as challenging testbeds, as they consist of real-life objects (*i.e.*, shoes, toys) and backgrounds (*i.e.*, sky, grass, marble floor) captured from large camera motions.

Waymo Open Dataset [34] consists of high-resolution video sequences recorded at a resolution of 1280×1920 using a multi-camera system installed on Waymo vehicles. It includes 798 training scenes and 202 validation scenes, each lasting 20 seconds and captured at 10 frames per second.

4.1.2. Baselines

G-SWM [21] is one of the representative models in the domain of dynamics prediction from images, placing particular emphasis on object-centric representations. It models object interactions using appearance information through a graph neural network and employs hierarchical latent modeling to capture temporal dynamics over time.

SlotFormer [44] (current **SOTA**) is designed for unsupervised visual dynamics simulation with object-centric representations. It leverages the Slot Attention framework to extract a set of slots that encode object attributes and their spatial relationships within a video frame.

OCVP [37] is an extension of SlotFormer that further explores the transformers by separating the attention block into specialized temporal and relational attention blocks. OCVP-Seq refers to the sequential processing of temporal and relational attention blocks, while OCVP-Par refers to the parallel processing of the two attention blocks.

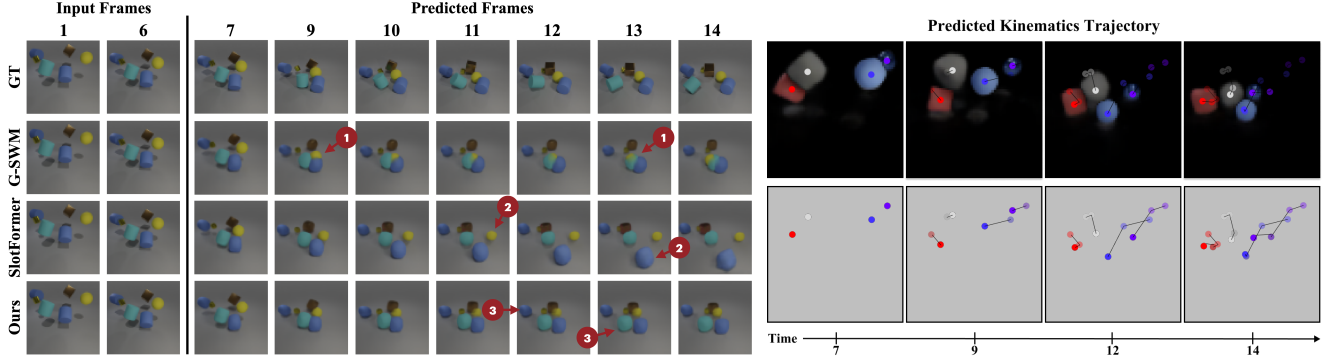


Figure 4. Generation results on MOVi-A. (a) We display the ground-truth frames (top row) and prediction results from two baseline models and Cross-OCK for comparison. (b) We visualize the kinematics trajectories of foreground objects across predicted frames. The top row displays the predicted kinematic trajectory overlaid on the frame, and the bottom row shows the isolated trajectory. Darker dots denote object positions in the latest predicted frame, while lighter dots, with progressively fading colors, represent earlier predictions.

Model	OBJ3D			MOVi-A			MOVi-B			MOVi-C			MOVi-D			MOVi-E		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
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	<u>0.932</u>	<u>0.024</u>	25.180	0.785	0.134	21.329	0.690	0.215	19.482	0.456	0.534	20.675	<u>0.565</u>	0.332	21.269	0.547	<u>0.335</u>
OCVP-Seq	33.100	<u>0.932</u>	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	<u>21.909</u>	0.688	0.226	17.941	0.402	0.650	Diverge			Diverge		
Joint-OCK	35.125	0.958	0.019	<u>27.259</u>	<u>0.811</u>	<u>0.124</u>	21.646	0.695	0.198	<u>21.038</u>	0.593	0.370	<u>22.087</u>	0.557	<u>0.282</u>	<u>22.394</u>	<u>0.569</u>	0.302
Cross-OCK	<u>34.097</u>	0.925	0.019	27.576	0.812	0.123	21.482	0.703	<u>0.209</u>	21.040	<u>0.592</u>	<u>0.376</u>	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.

Model	MOVi-A		MOVi-B		MOVi-C		MOVi-D		MOVi-E	
	FG-ARI \uparrow	mIOU \uparrow	FG-ARI \uparrow	mIOU \uparrow	FG-ARI \uparrow	mIOU \uparrow	FG-ARI \uparrow	mIOU \uparrow	FG-ARI \uparrow	mIOU \uparrow
G-SWM	0.431	0.500	0.410	0.443	0.238	0.414	0.239	<u>0.327</u>	0.374	<u>0.012</u>
SlotFormer	0.452	0.505	0.398	0.444	0.240	0.415	0.235	0.325	0.368	0.011
OCVP-Seq	0.433	0.503	0.403	0.448	0.121	0.410	Diverge		Diverge	
OCVP-Par	0.435	0.501	0.364	0.449	0.113	0.402	Diverge		Diverge	
Joint-OCK	<u>0.560</u>	<u>0.541</u>	0.483	0.453	0.347	0.528	<u>0.429</u>	0.482	<u>0.379</u>	0.019
Cross-OCK	0.563	0.547	0.481	<u>0.452</u>	<u>0.339</u>	0.515	0.430	0.482	0.380	0.020

Table 2. Evaluation of unsupervised video decomposition across five MOVi datasets. We exclude OBJ3D due to the absence of segmentation masks, which are necessary for generating the decomposition of individual slots.

4.1.3. Evaluation metrics

Video Prediction. To assess the visual quality of the predicted videos, we report PSNR [15], SSIM [41], and LPIPS [47], where LPIPS demonstrates the highest perceptual similarity to human perception compared to PSNR and SSIM. Despite the limitations of PSNR and SSIM in accurately assessing video quality [27], our model demonstrates strong performance in all metrics.

Scene Decomposition. We calculate the foreground Adjusted Rand Index (FG-ARI) and mean Intersection over Union (mIoU) to evaluate the predicted object dynamics us-

ing per-slot object masks generated by the SAVi decoder. FG-ARI is a similarity metric that assesses the correspondence between predicted and ground-truth segmentation masks in a permutation-invariant manner, renowned for its comprehensive evaluation of the predicted dynamics [20].

4.2. Video prediction on synthetic dataset

OCK demonstrates robust performance across environments of varying complexity, showcasing its ability to generate perceptually realistic predictions closely aligned with human perception as presented in Tab. 1. Notably, Cross-OCK achieves superior performance on complex MOVi-

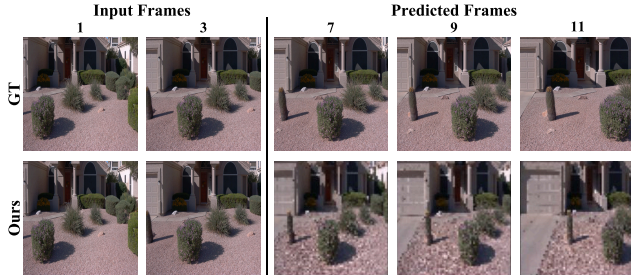


Figure 5. Generation results of OCK on Waymo Open Dataset.

Model	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
SlotFormer	19.127	0.330	0.714
OCVP-Seq	18.983	0.329	0.718
OCVP-Par		Diverge	
Joint-OCK	<u>25.023</u>	0.798	<u>0.251</u>
Cross-OCK	25.979	<u>0.728</u>	0.220

Table 3. Evaluation of video prediction on Waymo Open dataset against transformer-based baseline models.

{D,E} datasets by enabling the model to focus on relevant features specific to individual objects. We present qualitative results in Fig. 4a, where our model successfully generates future frames despite minor discrepancies (marked ③) in certain object locations compared to the ground-truth frames. These subtle differences have minimal impact on the overall prediction quality. In contrast, baseline models exhibit significant shortcomings, such as objects collapsing into amorphous shapes (marked ①) or losing their appearance and generating incorrect dynamics (marked ②). Fig. 4b showcases the model’s ability to accurately extract kinematics features from a sequence of video frames, facilitating precise prediction of object interaction.

Our model consistently demonstrates the most reliable and visually plausible predictions. Both OCK models strategically prioritize key features by integrating high-level object embeddings with low-level visual cues, facilitating precise object slot extraction and underscoring the pivotal role of Object Kinematics in preserving coherent appearances and motions for accurate video prediction.

4.3. Scene decomposition on synthetic dataset

To evaluate the accuracy of the predicted frames, we assess the quality of per-slot video decomposition using segmentation masks obtained from the video prediction process. As reported in Tab. 2, OCK exhibits performance comparable to four baseline models and proves the validity of the visual quality of the predicted videos. Accordingly, this underscores the positive trend of our model surpassing well-established benchmarks. For qualitative per-slot decomposition results, please refer to the supplementary material.

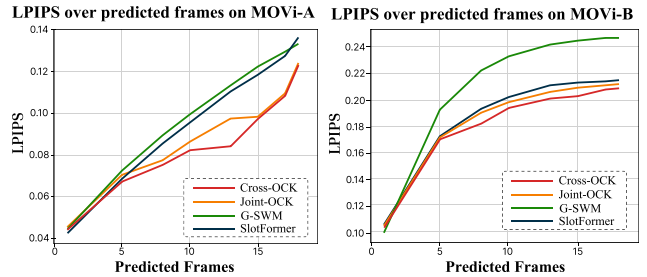


Figure 6. Long-term generalization results on the MOVi datasets, trained with 6 input frames and 8 future frames. Each video contains 24 frames, with generalization performance evaluated up to the 18th frame using the first 6 frames used as input.

4.4. Evaluation on real-world dataset

We evaluate video prediction performance in the Waymo Open Dataset in Tab. 3. We exclude scene decomposition due to the low reliability of the preprocessed segmentation masks. The Waymo Open Dataset, renowned for its visual complexity in real-world driving scenarios, serves as a challenging benchmark for object-centric learning. Consequently, high-speed vehicle motion and rapid background changes introduce inherent challenges, potentially affecting performance in real-world driving conditions. Despite these limitations, both OCK approaches achieve outstanding results compared to baseline models while preserving the visual quality of predicted videos. Moreover, as shown in Fig. 5, OCK successfully captures object interactions and makes structured predictions of object attributes and dynamics within the scene. For additional qualitative results, please refer to the supplementary material.

4.5. Evaluation on long-term generalization

To test long-term prediction capabilities, we evaluate OCK on two MOVi datasets over a longer horizon, surpassing the settings used in training. This is noteworthy as both datasets are trained with only six frames, significantly shorter than the full lengths. Even without additional regularization, OCK exhibits strong generalization for longer sequences during test time. In Fig. 6, the perceptual similarity of all models remains stable during initial frames but begins to diverge from the 7th predicted frame onward. Unlike baseline models that rely solely on object appearances and are prone to error accumulation, OCK effectively generalizes to variations in object motion and appearance, enhancing its long-term dynamic video prediction capability.

4.6. Ablation studies

In this section, we analyze how each transformer components influence the performance of dynamic video prediction in Tab. 4. We further conduct ablation studies on the two approaches, Joint-OCK and Cross-OCK, in Tab. 5.

Method	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Cross-OCK(A)	27.576	0.812	0.123
Joint-OCK(A)	27.259	0.811	0.124
Input Frame = 4	27.012	0.801	0.125
Input Frame = 8	27.122	0.806	0.125
Trans. Layer = 6	26.924	0.796	0.130
Trans. Layer = 8	26.503	0.784	0.133
Vanilla P.E.	23.600	0.591	0.205
Teacher Forcing	23.583	0.589	0.207

Table 4. Ablation study of transformer components on MOVi-A.

4.6.1. Transformer component analysis

Input frame. Our default training configuration for MOVi datasets involves utilizing six input frames and predicting eight frames. Increasing the number of input frames improves model performance; however, a slight decline is observed when further extending it to eight frames. This phenomenon arises from the adequacy of a six-frame history length for accurately capturing the object dynamics.

Transformer (Trans.) layer. We utilize four transformer layers as the default for MOVi-A. Additional layers result in unstable model training, and further augmentation causes diverging loss metrics. This underscores the crucial balance between the depth of the transformer model and training stability in addressing the complexities of object dynamics.

Vanilla positional encoding (P.E.). Our work incorporates temporal positional encoding such that object slots at the same timestep receive identical positional encoding. To evaluate its significance, we compare our approach with sinusoidal positional encoding, which disrupts permutation equivariance among object slots. As shown in Tab. 4, preserving permutation equivariance serves as a crucial prior, thus, maintaining the equivariance is essential to ensure accurate long-term dynamic video prediction.

Teacher forcing. We replace our model to incorporate a teacher forcing strategy [26] by feeding ground-truth object slots instead of predicted object slots during training. Surprisingly, this leads to a significant decline in performance, particularly affecting performance over longer prediction horizons. Thus, it highlights the importance of learning to handle its imperfect predictions for accurate dynamics modeling during training to optimize real-world predictions.

4.6.2. Object kinematics analysis

We evaluate both analytical and empirical approaches to investigate the potential impact of Object Kinematics computation variations on overall performance in Tab. 5. We train our model using six frames to predict eight subsequent frames and report the results for predicting 10 frames. Our analysis reveals a slight performance advantage for the analytical approach over the empirical approach. This is attributed to the fact that using the temporal positional state

Method	MOVi-A			MOVi-B		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Cross-OCK(A)	27.576	0.812	0.123	21.482	0.703	0.209
Joint-OCK(A)	27.259	0.811	0.124	21.494	0.695	0.212
Cross-OCK(E)	27.536	0.791	0.125	21.480	0.693	0.213
Joint-OCK(E)	26.750	0.801	0.125	21.482	0.694	0.213
Transformer	25.180	0.785	0.134	21.329	0.690	0.215

Table 5. Kinematic ablation of both transformer mechanisms on the MOVi-{A,B} datasets. (A) for the analytical approach, and (E) denotes the empirical approach.

of the current frame for guidance can be advantageous, but object dynamics in complex environments are often aperiodic, such that an inductive gradient-based learning may introduce confusion. Therefore, the analytical approach, which calculates the anticipated Object Kinematics of the subsequent video frame and integrates it with the current frame’s kinematics, generates more accurate frames. This is especially significant in complex dynamic environments where objects collide and consistently change their appearance since it enables the model to predict accurate dynamics and generate future frames with even greater precision.

5. Limitations

Despite considerable advancement in realism compared to previous work, we find evidence of a significant gap in the current slot-based object-centric encoder, which hinders its scalability in in-the-wild datasets [28]. Our framework can easily integrate state-of-the-art pretrained slot encoders, thus, a possible solution is to substitute the slot encoder to extract object representations from real-life videos, which would be a significant contribution to future work.

6. Conclusion

In this paper, we present OCK, an object-centric video prediction model that captures intricate object details by leveraging time-static object slots and time-varying Object Kinematics in an unsupervised manner. We propose a novel component named Object Kinematics along with two transformer architectures to enhance dynamic video prediction performance in complex synthetic and real-world datasets. Experimental results demonstrate that our model effectively captures spatiotemporal object patterns by utilizing time-varying kinematic attributes, surpassing prior methods that primarily emphasize static object appearances and explicit scene-level interactions. Incorporating time-varying kinematic information shows significant potential for modeling object dynamics, enabling improved object-centric video prediction and generation in real-world environments.

Acknowledgements

This work was partly supported by the IITP (RS-2021-II212068-AIHub/10%, RS-2021-II211343-GSAI/15%, RS-2022-II220951-LBA/15%, RS-2022-II220953-PICA/20%), NRF (RS-2024-00353991-SPARC/20%, RS-2023-00274280-HEI/10%), and KEIT (RS-2024-00423940/10%) grant funded by the Korean government.

References

- [1] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34:20014–20027, 2021. 5
- [2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2
- [3] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019. 2
- [4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2
- [5] Ayush Chakravarthy, Trang Nguyen, Anirudh Goyal, Yoshua Bengio, and Michael C Mozer. Spotlight attention: Robust object-centric learning with a spatial locality prior. *arXiv preprint arXiv:2305.19550*, 2023. 2
- [6] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 4
- [7] Tal Daniel and Aviv Tamar. Ddip: Unsupervised object-centric video prediction with deep dynamic latent particles. *arXiv preprint arXiv:2306.05957*, 2023. 1
- [8] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D Hwang, et al. Faith and fate: Limits of transformers on compositionality (2023). *arXiv preprint arXiv:2305.18654*, 2023. 3
- [9] Martin Engelcke, Adam R Kosiorek, Oiwi Parker Jones, and Ingmar Posner. Genesis: Generative scene inference and sampling with object-centric latent representations. *arXiv preprint arXiv:1907.13052*, 2019. 2
- [10] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3170–3180, 2022. 2
- [11] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *International Conference on Machine Learning*, pages 2424–2433. PMLR, 2019. 2
- [12] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanaprasgam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3749–3761, 2022. 5
- [13] Kyle Hamilton, Aparna Nayak, Bojan Božic, and Luca Longo. Is neuro-symbolic ai meeting its promise in natural language processing? a structured. *arXiv preprint arXiv:2202.12205*, 2022. 1
- [14] Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, and Shuchang Zhou. A dynamic multi-scale voxel flow network for video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6121–6131, 2023. 4
- [15] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008. 6
- [16] Jindong Jiang and Sungjin Ahn. Generative neurosymbolic machines. *Advances in Neural Information Processing Systems*, 33:12572–12582, 2020. 1
- [17] Daniel Kahneman, Anne Treisman, and Brian J Gibbs. The reviewing of object files: Object-specific integration of information. *Cognitive psychology*, 24(2):175–219, 1992. 1
- [18] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy J Mitra. Holodiffusion: Training a 3d diffusion model using 2d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18423–18433, 2023. 4
- [19] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in Neural Information Processing Systems*, 31, 2018. 2
- [20] Thomas Kipf, Gamaleldin F Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. *arXiv preprint arXiv:2111.12594*, 2021. 5, 6
- [21] Zhixuan Lin, Yi-Fu Wu, Skand Peri, Bofeng Fu, Jindong Jiang, and Sungjin Ahn. Improving generative imagination in object-centric world models, 2020. 1, 2, 3, 5
- [22] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in Neural In-*

- formation Processing Systems*, 33:11525–11538, 2020. 2, 3
- [23] Ruibo Ming, Zhewei Huang, Zhuoxuan Ju, Jianming Hu, Lihui Peng, and Shuchang Zhou. A survey on video prediction: From deterministic to generative approaches, 2024. 3
- [24] Michael Mohnhaupt and Bernd Neumann. Understanding object motion: Recognition, learning and spatiotemporal reasoning. *Robotics and Autonomous Systems*, 8(1-2):65–91, 1991. 3
- [25] Marc Oliu, Javier Selva, and Sergio Escalera. Folded recurrent neural networks for future video prediction. In *Proceedings of the European Conference on Computer Vision*, pages 716–731, 2018. 2
- [26] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training, 2018. 8
- [27] Umme Sara, Morium Akter, and Mohammad Shorif Uddin. Image quality assessment through fsim, ssim, mse and psnr—a comparative study. *Journal of Computer and Communications*, 7(3):8–18, 2019. 6
- [28] Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. *arXiv preprint arXiv:2209.14860*, 2022. 8
- [29] Javier Selva, Anders S Johansen, Sergio Escalera, Kamal Nasrollahi, Thomas B Moeslund, and Albert Clapés. Video transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3
- [30] Gautam Singh, Fei Deng, and Sungjin Ahn. Illiterate dall-e learns to compose. *arXiv preprint arXiv:2110.11405*, 2021. 2
- [31] Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. In *Advances in Neural Information Processing Systems*, 2022. 1, 2
- [32] Yeon-Ji Song, Hyunseo Kim, Suhyung Choi, Jin-Hwa Kim, and Byoung-Tak Zhang. Learning object motion and appearance dynamics with object-centric representations. In *Causal Representation Learning Workshop at NeurIPS*, 2023. 2
- [33] Yeon-Ji Song, Jaemin Kim, Byung-Ju Kim, and Byoung-Tak Zhang. Dbmovi-gs: Dynamic view synthesis from blurry monocular video via sparse-controlled gaussian splatting, 2025. 2
- [34] Pei Sun, Henrik Kretschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 5
- [35] Qu Tang, Xiangyu Zhu, Zhen Lei, and Zhaoxiang Zhang. Intrinsic physical concepts discovery with object-centric predictive models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23252–23261, 2023. 1, 2
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3
- [37] Angel Villar-Corrales, Ismail Wahdan, and Sven Behnke. Object-centric video prediction via decoupling of object dynamics and interactions. In *International Conference on Image Processing (ICIP)*, 2023. 2, 5
- [38] Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li Fei-Fei. Eidetic 3d lstm: A model for video prediction and beyond. In *International conference on learning representations*, 2018. 2
- [39] Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, S Yu Philip, and Mingsheng Long. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2208–2225, 2022. 2
- [40] Yanbo Wang, Letao Liu, and Justin Dauwels. Slot-vae: Object-centric scene generation with slot attention. *arXiv preprint arXiv:2306.06997*, 2023. 2
- [41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6
- [42] Sean Welleck, Peter West, Jize Cao, and Yejin Choi. Symbolic brittleness in sequence models: on systematic generalization in symbolic mathematics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8629–8637, 2022. 3
- [43] Yue Wu, Qiang Wen, and Qifeng Chen. Optimizing video prediction via video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17814–17823, 2022. 2
- [44] Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models. *arXiv preprint arXiv:2210.05861*, 2022. 1, 2, 3, 4, 5
- [45] Andrii Zadaianchuk, Maximilian Seitzer, and Georg Martius. Self-supervised visual reinforcement learning with object-centric representations. In *International Conference on Learning Representations*, 2021. 1
- [46] Andrii Zadaianchuk, Georg Martius, and Fanny Yang. Self-supervised reinforcement learning with independently controllable subgoals. In *Conference on Robot Learning*, pages 384–394. PMLR, 2022. 1
- [47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6
- [48] Zixu Zhao, Jiase Wang, Max Horn, Yizhuo Ding, Tong He, Zechen Bai, Dominik Zietlow, Carl-Johann Simon-Gabriel, Bing Shuai, Zhuowen Tu, et al. Object-centric multiple object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16601–16611, 2023. 1