Learning Object Motion and Appearance Dynamics with Object-Centric Representations

Yeon-Ji Song[1] **Hyunseo Kim**[1] **Suhyung Choi**[1]
Jin-Hwa Kim[2,3*] **Byoung-Tak Zhang**[1,2*]
[1]Seoul National University [2]AIIS [3]NAVER AI Lab
{yjsong, hskim, schoi}@bi.snu.ac.kr
j1nhwa.kim@navercorp.com btzhang@bi.snu.ac.kr

Abstract

Human perception involves discerning objects based on attributes such as size, color, and texture, and making predictions about their movements using features such as weight and speed. This innate ability operates without the need for conscious learning, allowing individuals to perform actions like catching or avoiding objects when they are unaware. Accordingly, the fundamental key to achieving higher-level cognition lies in the capability to break down intricate multi-object scenes into meaningful appearances. Object-centric representations have emerged as a promising tool for scene decomposition by providing useful abstractions. In this paper, we propose a novel approach to unsupervised video prediction leveraging object-centric representations. Our methodology introduces a two-component model consisting of a slot encoder for object-centric disentanglement and a feature extraction module for masked patches. These components are integrated through a cross-attention mechanism, allowing for comprehensive spatio-temporal reasoning. Our model exhibits better performance when dealing with intricate scenes characterized by a wide range of object attributes and dynamic movements. Moreover, our approach demonstrates scalability across diverse synthetic environments, thereby showcasing its potential for widespread utilization in vision-related tasks.

1 Introduction

Human-level intelligent system requires an understanding of the environments and spatio-temporal interactions of the surrounding objects [6]. This proficiency holds paramount importance as it underpins the human ability to perceive visual scenes as a geometric composition of objects' components, thereby recognizing objects in a dynamically changing world [10]. Recently, object-centric learning has emerged with the aim of modeling the compositional structure of a scene into a set of object representations. The advent of object-centric learning has demonstrated value in scenarios where the set of objects encountered may be diverse and constantly changing, or even not previously encountered. Accordingly, it has presented remarkable performance across a wide range of domains such as unsupervised scene understanding [9, 20, 21], object tracking [8, 26], and reinforcement learning [29, 30]. Particularly, object-centric representations have shown promising results in video prediction tasks, as they not only predict future frames but also learn spatio-temporal dynamics of object interactions on the basis of their motion and appearance.

Despite the recent success of object-centric models for processing spatio-temporal dynamics, effectively capturing the intricate properties of objects (i.e., appearance and motion) in both simple and synthetic environments remains a significant challenge [13, 26]. This challenge arises not only

*Corresponding authors.

Accepted to the NeurIPS 2023 Workshop on Causal Representation Learning.

from the object-centric learning mechanism itself but also from how it is employed and executed in downstream tasks. For instance, previous approaches have faced limitations in training on large image and video datasets due to architectural complexities or have found object-centric models to be insufficient in extracting crucial information for object-related tasks [12, 21, 26]. Overcoming these limitations is crucial, as it allows us to leverage extensive amounts of images and videos, leading to stronger spatio-temporal interactions between objects. This, in turn, enhances scalability and promotes generalization to complex environments.

In this paper, we propose a method called Object-centric Slot Patch Transformer, which is a dynamics learning mechanism that predicts future frames in an object-centric way. Specifically, we employ cross-attention based on cross-covariance attention [1] to fuse features extracted from two distinct components: (1) Slot Attention module, which disentangles scenes into a set of object-centric representations (a.k.a. slots), and (2) Masked patch extraction module, in which we randomly mask patches of the scene. The cross-attention mechanism that combines two separate embeddings significantly advances our understanding of the spatio-temporal arrangement and relationships between objects in the scene. Furthermore, our proposed cross-attention mechanism with object-centric representations notably reduces the overall computation of the attention layer in the transformer, thereby enhancing the efficiency of the model, particularly in tasks relying on large visual data. We empirically demonstrate the effectiveness of our proposed method in both simple and synthetic environments, and underscore its feasibility under complex environments, such as MOVi datasets [5], involving various object appearances and motions (i.e., collision, attraction, or repulsion). Notably, our proposed method achieves comparable performance to existing approaches in both simple and complex environments.

2 Related work

Unsupervised video prediction for dynamics modeling. Unsupervised video prediction is a task that predicts future frames or sequences with the absence of auxiliary annotations [4]. This task leverages the inherent temporal consistency in video data to learn meaningful representations of visual appearances, interactions between objects, and dynamic changes in the scene. Earlier methods utilize encoder-decoder architectures, with recurrent or transformer-based modules, to model the underlying spatio-temporal relationships [16, 22, 24]. Other works adapt transformer models for video prediction [26, 27], but the challenges lie in designing robust models that handle complex scenes and diverse motions, as they lack understanding of objects' intrinsic properties. In contrast to the previous works, our method performs unsupervised object-centric video prediction in complex environments by modeling both extrinsic and intrinsic properties of objects for a higher-level understanding of video scenes.

Object-centric representation learning. Recent work on unsupervised object-centric representation learning decomposes visual scenes into a set of vectors named *slots*, which are a set of representations that capture the compositional properties of visual scenes, via a mechanism named Slot Attention [14]. Slot Attention has shown remarkable performance in diverse vision-based tasks such as object discovery [11, 19], scene reconstruction [13, 20, 21], dynamics learning [26], or even solving certain types of RL problems [29]. Among these, our work focuses on dynamics learning by treating it as a video prediction problem and applies Slot Attention to extract object-centric representations from the video. While existing work on unsupervised video prediction with object-centric representations models long-term interactions, it has shown limitations in handling complex environments, where an adoption of traditional transformers appears to be a contributing factor [26].

Cross-covariance attention. Transformers have gained recognition for their effectiveness in handling sequential data [23]. The self-attention mechanism in transformers enables precise localization and segmentation of objects in static scenes. Accordingly, transformers in dynamics modeling, with a particular emphasis on object-centric representations, predict objects' motion in simple visual scenes over a sequence of frames [7, 21]. However, the self-attention mechanism has shown limitations in processing dense visual data, particularly in object-centric scene decomposition and prediction [12, 26]. In order to reduce the complexity of the self-attention mechanism, Ali et al. [1] introduced Cross-Covariance Attention mechanism, which replaces the self-attention layer with a cross-covariance attention block. Our method performs unsupervised object-centric dynamics modeling with the aid of cross-covariance attention, which operates across feature channels rather than tokens, allowing efficient processing of large datasets (i.e., high-resolution images or videos).

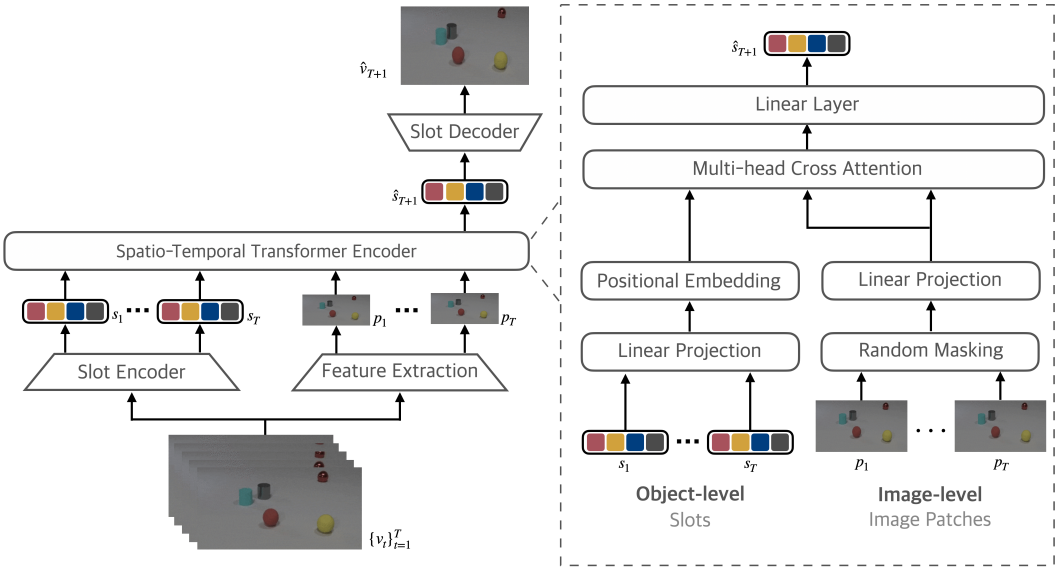

Figure 1: Main architecture overview. Our method computes cross-attention (dotted box) between slots (left part in the dotted box) and masked patches (right part in the dotted box) to acquire representations of future frames.

3 Object-centric Slot Patch Transformer

In this section, we propose the autoregressive dynamics prediction model that captures the spatial and temporal relationships between objects, which we call the Object-centric Slot Patch Transformer. As illustrated in Figure 1, our method performs cross-attention between two components: slots and image patches, described in Section 3.1 and Section 3.2, respectively. The two features are passed through the cross-attention mechanism along with slots, which enables a deeper spatio-temporal understanding of objects and its environment, as explained in Section 3.3. Finally, Section 3.4 describes the overall autoregressive training procedure, as we re-use the prediction output as an input in the next training step.

3.1 Slot-based object representation

We build on the Slot Attention [14] to extract slots from video sequences. Given a set of T video frames $\{v_t\}_{t=1}^{T}$, we employ Convolutional Neural Network (CNN) [18] encoder to extract image features. These embedded features are enriched with positional encoding. The resulting embeddings, which contain content and positional information of the scene, are then flattened into a set of vectors $h_t = \mathbb{R}^{N_{slot} \times D_{\text{slot}}}$, where N_{slot} is the size of the flattened feature and D_{slot} is the CNN feature dimension. Then our object-centric module initializes slots based on the set of vectors h_t, and performs Slot Attention f_{SA} to update the set of slot representations $S_t = \{s_1, ..., s_N\}$ via iterative Scaled Dot Product Attention.

3.2 Masked patch extraction

Simultaneously, we conduct feature extraction that outputs masked patches, which are fed into the cross-attention mechanism along with the extracted slots. The intuition behind this approach is to preserve the properties of representations so that the model does not learn incorrect dynamics during the initial stages of training, which is a phenomenon often observed when solely relying on the Slot Encoder. Specifically, the incorporation of masked patches guides the model to focus on specific regions, preventing it from prematurely learning unreliable patterns. This focused attention, combined with the extracted slots, enables the model to capture nuanced information and relationships within the visual observation.

Given a set of T video frames $\{v_t\}_{t=1}^{T}$, we apply a patch-aligned random masking strategy for each frame. We split each frame $x \in \mathbb{R}^{H \times W \times C}$, where (H, W) is the input image resolution and C corresponds to the number of channels, into a sequence of non-overlapping patches $\{x_n^p\}_{n=1}^{N_{patch}}$ with a size of $P \times P$. The image patches of dimension D_{patch} are then linearly projected $p_t \in \mathbb{R}^{(P^2C) \times D_{\text{patch}}}$ and randomly masked with a hyper-parameter λ, which are then replaced with a learnable embedding. This results in a final output of masked image patches P_t, which are linearly projected to match the dimension of the cross-attention transformer.

3.3 Cross-attention mechanism

Our method computes in a form of cross attention [2, 15] between extracted slots from the Slot Attention encoder, as `queries`, and masked patch embeddings as `keys` and `values`.

Priorly, we perform pretraining on the Slot Attention model using reconstruction loss on videos from the target data. This pretraining process attains enhanced accuracy and comprehensiveness of the object-centric representation as it ensures that the learned slots adeptly capture the visual characteristics of both foreground objects and background objects within the scene.

We then compute the cross-attention mechanism. First, we linearly project input sequences of slots and masked patches to a set of latent space, $S_t = \textit{Linear}(s_t)$ and $P_t = \textit{Linear}(p_t)$, respectively. Following SlotFormer Wu et al. [26], we apply positional embeddings at a temporal level to maintain the frame order and permutation equivariance. Then, cross-attention transformer operates along the latent slots $S_t \in \mathbb{R}^{N_{\text{slot}} \times D_{\text{slot}}}$ over masked patch tokens $P_t \in \mathbb{R}^{N_{\text{patch}} \times D_{\text{patch}}}$. In detail, each slot representation is updated by performing cross-attention of its original slot representation over the attended value representation obtained from the masked patch sequence. Unlike the traditional transformer, where we scale down the weights by the dimensions of key vectors, inspired by cross-covariance attention [1], we use a temperature parameter τ that adjusts the inner products prior to applying the Softmax operation.

Mathematically, the cross-attention(`CA`) mechanism can be described as:

$$\mathbf{q} = S_t \mathbf{W}_q, \quad \mathbf{k} = P_t \mathbf{W}_k, \quad \mathbf{v} = P_t \mathbf{W}_v, \tag{1}$$

$$\mathbf{A} = \texttt{softmax}(\frac{\mathbf{k}^\top \cdot \mathbf{q}}{\tau}), \quad \texttt{CA}(s_t) = \mathbf{A} \cdot \mathbf{v}^\top \tag{2}$$

The implementation of the cross-attention module yields object-centric representations that are effective and contextually meaningful. Consequently, the model demonstrates proficiency in capturing dependencies and interrelationships among objects.

3.4 Model training

Our model is trained by taking the last N output features of the transformer $\mathcal{T}$ and feeding them to a linear layer to update the slots at the next timestep $\hat{S}_{T+1}$. For consequence future predictions, $\hat{S}_{T+1}$ is fed as input along with the ground-truth slots, and the transformer is applied to generate and predict longer frames in an autoregressive manner. As our model aims to predict future frames, we train to minimize slot reconstruction loss and image reconstruction loss as follows:

$$\mathcal{L} = \mathcal{L}_{\text{slot}} + \lambda \mathcal{L}_{\text{image}} \tag{3}$$

The slot reconstruction loss $\mathcal{L}_{\text{slot}}$ is the L2 loss between the ground-truth slot S_{T+1} and the reconstructed slot $\hat{S}_{T+1}$. For image reconstruction loss $\mathcal{L}_{\text{image}}$, we use frozen SAVi [11] decoder f_θ^{savi} to convert predicted slots $\hat{S}_{T+k}$ to image, and then match with the ground-truth image x_{T+k}.

Empirical evidence supporting this claim is presented in Section 4, where we experimentally show that cross-attention outperforms alternative approaches, particularly on complex scenes.

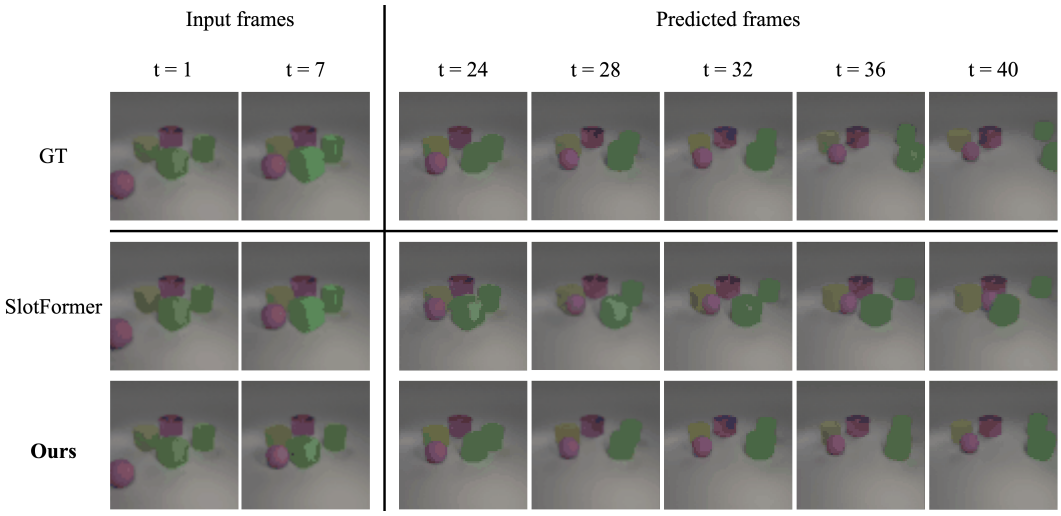

Figure 2: Generation results on OBJ3D. We train the baseline model (top) and our model (bottom) with 6 frames and generate 10 frames, which is claimed sufficient for the model to learn accurate dynamics as described in Wu et al. [26].

		G-SWM		SlotFormer		Ours	
Dataset	**Complexity**	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑
OBJ3D	simple	0.10	31.43	0.08	32.26	**0.06**	**32.82**
CLEVRER	simple	0.16	28.42	0.17	31.34	**0.14**	**32.38**
MOVi-A	complex	0.46	23.57	0.30	**25.42**	**0.28**	24.46
MOVi-C	complex	0.69	15.50	0.66	19.34	**0.42**	**20.97**

Table 1: Evaluation of visual quality on the datasets, ordered based on level of difficulty. Details of the datasets with figures in Appendix A.

4 Experiments

4.1 Implementation details

Models and datasets. We compare our method with two baseline models: G-SWM [12] and SlotFormer [26]. Both of these models excel in object reasoning tasks, leveraging object-centric representations to gain profound insights into intricate visual scenes. These approaches enable the model to process and reason about their attributes and interactions in a structured manner. We evaluate our approach on five different datasets, encompassing a range of complexities from simple to synthetic video datasets: OBJ3D [12], CLEVRER [28], and a variety of MOVi datasets implemented using Kubric [5]. Further details are provided in Appendix A.

Metrics. We compare the visual quality of generated videos using PSNR [25] and LPIPS [31]. As highlighted in Zhang et al. [32] and Sara et al. [17], PSNR and SSIM metrics do not closely correspond to human perception. In contrast, LPIPS exhibits a more reliable correlation with human perceptual judgments, leveraging learned deep features. Consequently, our comparative analysis predominantly centers around the LPIPS metric, and PSNR for reference only.

4.2 Video prediction

Table 1 summarizes the generation results of our model and the baseline models. While SlotFormer exhibits the capacity to generate future frames in the OBJ3D and CLEVRER datasets, our approach outperforms the baselines in terms of LPIPS scores and achieves competitive results in terms of PSNR.Notably, while our model yields a similar PSNR score to the baseline model for OBJ3D, Figure 2 illustrates that it predicts future frames with a consistent appearance and motion compared

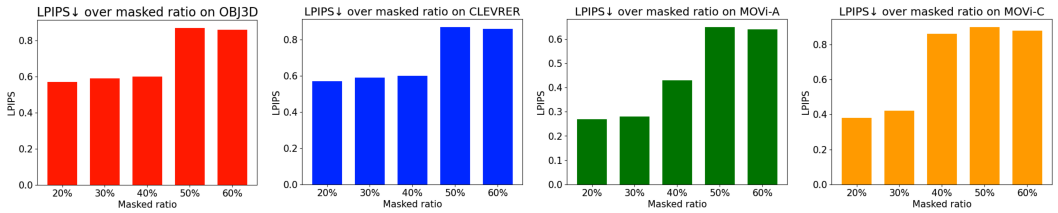

Figure 3: Evaluation on masking ratios with our model across all datasets using LPIPS metric.

to the baseline model. For example in SlotFormer, starting from $t = 36$, the pink ball moves in an incorrect direction, despite an earlier observation of it colliding with the yellow cube around $t = 30$. In contrast, our model demonstrates consistent prediction of object appearance and motion over time. This observation underscores that PSNR may not be the most robust metric for assessing generation quality. In the case of MOVi-A and MOVi-C, our model generates consistent frames and achieves superior results compared to both baseline models. This underscores the robustness of our approach in handling complex environments characterized by changes in object appearance and motion over time.

4.3 Detailed analyses

Image reconstruction with cross-attention mechanism. We further evaluate the capability of image decomposition and reconstruction ability on OBJ3D to verify that the integration of slots from the slot encoder and masked patches from the feature extraction module, facilitated by the cross-attention mechanism, does not impede the unsupervised video prediction task. Our findings confirm that the cross-attention mechanism successfully segments objects within the scene. Generation results of image reconstruction and decomposition on OBJ3D are provided in Appendix B.

Training sequence length. In our study, SlotFormer serves as the reference for determining experimental setting across OBJ3D and CLEVRER datasets. In the case of MOVi-A and MOVi-C, we conducted training with a carefully chosen set of hyperparameters, taking into consideration both computational resources and the setup implemented in SlotFormer. For a detailed and comprehensive overview of the specific hyperparameters used for each dataset, we direct readers to Appendix A.

Masked image patches. We conducted a performance evaluation of our model across various masking ratios on each dataset, spanning from 20% to 60%. Masking ratios below 20% yielded results nearly identical to the original image, while ratios above 60% rendered meaningless patches that masked out a majority of the important object components. As shown in Figure 3, we evaluated based on LPIPS score. We observed a notable improvement in performance at a specific ratio for each dataset, and thus, opted to use the particular ratios for our training.

5 Conclusion and future work

In this paper, we propose a new dynamics prediction mechanism with object-centric representations. Our model captures intricate relationships between objects, encompassing both appearance and motion characteristics, in an unsupervised way. Experimental results demonstrate that our model predicts better object appearance and motion compared to the baselines in complex environments. This distinction arises from our model's ability to comprehend spatial and temporal patterns of objects facilitated by two components we propose, as opposed to existing architectures that primarily focus on modeling explicit scene-level object interactions. Additionally, our approach introduces a cross-attention mechanism with object-centric representations, which operates over a set of feature vectors for each image embedding. This significantly reduces computation demands of the traditional self-attention mechanism, thereby accelerating training speed, making it well-suited for tasks involving dynamics prediction and video generation.

A future improvement of this work could be to enhancing its potential for longer dynamics prediction and generalization to diverse datasets. These advancements hold significant promise in expanding our model to various downstream tasks, including visual question answering, action planning, and even in the domain of autonomous navigation.

Acknowledgments and Disclosure of Funding

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (2021-0-02068-AIHub/15%, 2021-0-01343-GSAI/20%, 2022-0-00951-LBA/25%, 2022-0-00953-PICA/25%) and NRF (RS-2023-00274280/15%) 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.

[2] 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.

[3] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022.

[4] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction, 2016.

[5] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, 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, June 2022.

[6] Kyle Hamilton, Aparna Nayak, Bojan Božić, and Luca Longo. Is neuro-symbolic ai meeting its promises in natural language processing? a structured review. *Semantic Web*, (Preprint):1–42, 2022.

[7] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3148–3159, 2022.

[8] Jindong Jiang and Sungjin Ahn. Generative neurosymbolic machines. *Advances in Neural Information Processing Systems*, 33:12572–12582, 2020.

[9] Jindong Jiang, Sepehr Janghorbani, Gerard De Melo, and Sungjin Ahn. Scalor: Generative world models with scalable object representations. *arXiv preprint arXiv:1910.02384*, 2019.

[10] 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.

[11] 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.

[12] Zhixuan Lin, Yi-Fu Wu, Skand Peri, Bofeng Fu, Jindong Jiang, and Sungjin Ahn. Improving generative imagination in object-centric world models, 2020.

[13] Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. Space: Unsupervised object-oriented scene representation via spatial attention and decomposition. *arXiv preprint arXiv:2001.02407*, 2020.

[14] 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 Information Processing Systems*, 33:11525–11538, 2020.

[15] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[16] MarcAurelio Ranzato, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*, 2014.

[17] 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.

[18] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61: 85–117, 2015.

[19] 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.

[20] Gautam Singh, Fei Deng, and Sungjin Ahn. Illiterate DALL-e learns to compose. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=h0OYV0We3oh`.

[21] 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.

[22] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852. PMLR, 2015.

[23] 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.

[24] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033*, 2017.

[25] 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.

[26] 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.

[27] Xi Ye and Guillaume-Alexandre Bilodeau. Vptr: Efficient transformers for video prediction, 2022.

[28] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

[29] Andrii Zadaianchuk, Maximilian Seitzer, and Georg Martius. Self-supervised visual reinforcement learning with object-centric representations. In *International Conference on Learning Representations*, 2021.

[30] 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.

[31] 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.

[32] 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.

A Implementation details

A.1 Hyperparameters

	Dataset	OBJ3D	CLEVRER	MOVi-A	MOVi-C
Slot Encoder module	Training steps	80k	200k	150k	150k
	Base model	SAVi	SAVi	SAVi	SAVi
	Number of slots	7	8	11	11
	Slot size	128	128	128	128
	Batch size	64	64	32	32
	Input size	64×64	64×64	64×64	64×64
Patch Extraction module	Patch size	4×4	4×4	4×4	4×4
	Masking ratio	40%	40%	30%	30%
	Input size	64×64	64×64	64×64	64×64
Transformer module	Training steps	200k	400k	200k	200k
	Conditioned frames	6	6	6	6
	Predicted frames	10	10	10	10
	Batch size	64	64	32	32
	Learning rate	2×10^{-4}	2×10^{-4}	2×10^{-4}	2×10^{-4}
	Transformer layers	4	4	8	8
	Embedding size	128	256	256	256
	Dimension size	256	256	128	128
	Loss weight λ	1	1	0.7	0.7

Table 2: List of hyperparameters for each dataset.

A.2 Dataset details

(a) (b) (c) (d)

Figure 4: Overview of the datasets. (a) OBJ3D and (b) CLEVRER datasets are considered simple as the objects move on a flat surface. (c) MOVi-A, and (d) MOVi-C are challenging due to the diversity of objects' motion, where objects are tossed on the surface.

OBJ3D is crafted to address tasks related to the comprehension and manipulation of 3D objects [12]. It encompasses a diverse array of objects distinguished by their varying shapes, textures, and visual attributes, each accompanied by corresponding object-centric annotations. Each object is annotated with its 3D spatial orientation, geometric form, and semantic classification.

CLEVRER dataset incorporates temporal aspects by presenting videos of dynamic scenes wherein objects exhibit motion, collision, and interaction. It comprises artificially generated images composed of 3D objects, constructed using basic geometric shapes, diverse colors, materials, and spatial configurations. Annotations offer detailed information regarding object characteristics, interrelationships, and physical attributes, facilitating the comprehension and deduction of insights from intricate visual scenes. Notably, to enable an unsupervised evaluation of performance, the annotations are intentionally excluded from the training and prediction phases.

MOVi-A derived from the CLEVR dataset, encompasses a wider range of complex visual content. This dataset encompasses variations in lighting conditions, background complexity, occlusions, and

object appearances. The scenes in MOVi-A consist of 3 to 10 randomly positioned objects situated on a gray floor. The camera's position remains stationary, directed towards the origin point.

MOVi-C is comprised of genuine video sequences that displays moving objects situated within a range of backgrounds, characterized by diverse lighting conditions and contextual arrangements. In contrast to MOVi-A, which utilizes elementary shapes with uniform coloring, MOVi-C incorporates authentic objects obtained from the Google Scanned Objects (GSO) dataset [3]. The backgrounds are randomly generated from Poly Haven, which also serves as the ground surface.

A.3 Baseline details

G-SWM [12] is a leading-edge model in the domain of dynamics prediction from images, placing particular emphasis on object-centric representations. It excels in breaking down scenes into foreground and background components, while also disentangling object appearances. This unique capability allows G-SWM to model object interactions and dynamics. Through the adoption of object-centric representations, G-SWM acquires a nuanced understanding of how individual objects influence the overall dynamics, making it a powerful tool for tasks like video prediction and scene understanding.

SlotFormer [26] is a state-of-the-art model designed for dynamics prediction from images via object-centric representations. It employs Slot Attention, a mechanism that isolates and extracts a set of slots representing various components of a visual scene. These slots encapsulate information about object attributes and their spatial relationships. SlotFormer provides a comprehensive understanding of how objects interact and move within a given scene.

B Image decomposition and reconstruction result on OBJ3D

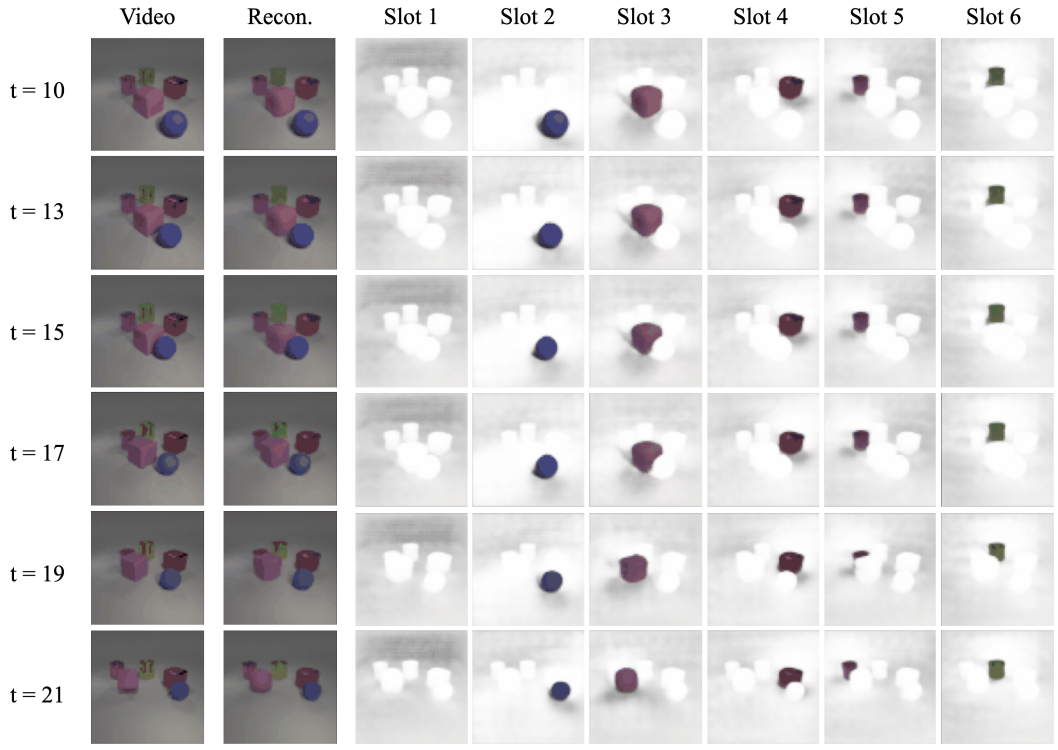

Figure 5: Unsupervised image decomposition and reconstruction performance on OBJ3D dataset trained on 6 slots. We show ground truth in the first column and the reconstructed image of the predicted slots in the subsequent columns.