Time-Conditioned Generative Modeling of Object-Centric Representations for Video Decomposition and Prediction

Chengmin Gao[1] **Bin Li**[*1]

[1]School of Computer Science, Fudan University

Abstract

When perceiving the world from multiple viewpoints, humans have the ability to reason about the complete objects in a compositional manner even when an object is completely occluded from certain viewpoints. Meanwhile, humans are able to imagine novel views after observing multiple viewpoints. Recent remarkable advances in multi-view object-centric learning still leaves some unresolved problems: 1) The shapes of partially or completely occluded objects can not be well reconstructed. 2) The novel viewpoint prediction depends on expensive viewpoint annotations rather than implicit rules in view representations. In this paper, we introduce a time-conditioned generative model for videos. To reconstruct the complete shape of an object accurately, we enhance the disentanglement between the latent representations of objects and views, where the latent representations of time-conditioned views are jointly inferred with a Transformer and then are input to a sequential extension of Slot Attention to learn object-centric representations. In addition, Gaussian processes are employed as priors of view latent variables for video generation and novel-view prediction without viewpoint annotations. Experiments on multiple datasets demonstrate that the proposed model can make object-centric video decomposition, reconstruct the complete shapes of occluded objects, and make novel-view predictions.

1 INTRODUCTION

Humans understand the multi-object world in a compositional manner that the representations of multiple objects are memorized separately and then combined into the perceived whole [Kahneman et al., 1992, Spelke and Kinzler, 2007, Johnson, 2010]. When it comes to the multi-object scene with multiple viewpoints, humans exhibit higher-level intelligence in multiple aspects: On one hand, a certain object is endowed with a canonical representation that depicts its complete 3D shape and appearance through multi-view perception [Turnbull et al., 1997]. As a result, humans have the ability to reason about the complete object even when an object is completely occluded from certain viewpoints [Shepard and Metzler, 1971]. On the other hand, scenes observed from novel viewpoints can be imagined on the basis of the learned implicit rules of perspective [Schacter et al., 2012, Beaty et al., 2016]. Such compositional modeling from multiple viewpoints is the fundamental ingredient for high-level cognitive intelligence.

Unsupervised object-centric learning that is dedicated to simulating human intelligence have recently achieved remarkable advances [Yuan et al., 2022a], especially in single-view object-centric learning on both images [Burgess et al., 2018, Yuan et al., 2019a,b, Engelcke et al., 2021] and videos [Kosiorek et al., 2018, Jiang et al., 2019, Lin et al., 2020]. Meanwhile, multi-view object-centric learning [Li et al., 2020, Chen et al., 2021, Kabra et al., 2021, Yuan et al., 2022b], which aims to learn 3D object representations, also demonstrates a promising blueprint; however, it still leaves some unresolved problems: 1) The shapes of partially or completely occluded objects from some viewpoints cannot be reconstructed through 3D representations learned from other viewpoints. Although some models can theoretically restore occlusions, relatively poor restoration (e.g. inaccurate shadows, blurs and noises) is inevitably observed. 2) Despite using the query objective during training [Li et al., 2020], the ability for novel viewpoint prediction depends on expensive viewpoint annotations, which provide strong location information and play a crucial role in update of object-centric representations; while the implicit rules of view representations are not fully explored to make prediction. It is, therefore, crucial to develop a unified multi-view model to perform object-centric learning like humans.

[*]Corresponding author (libin@fudan.edu.cn)

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

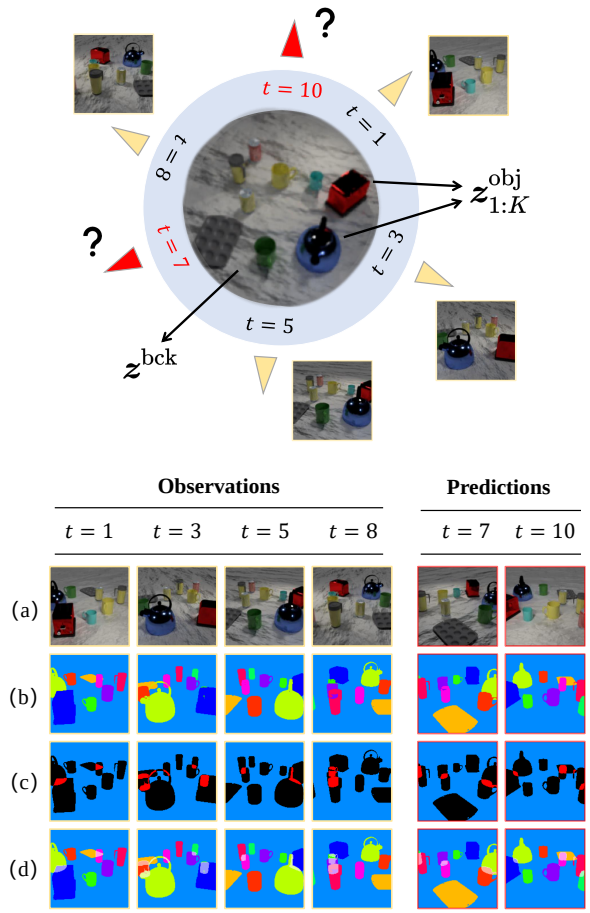

Figure 1: **Top**: Video decomposition and prediction with multiple observed time-conditioned viewpoints. The yellow and red triangles represent the observed frames and predicted frames, respectively. **Bottom**: The expected outputs: (a) reconstruction, (b) segmentation, (c) overlaps, and (d) complete segmentation. In our problem setting, only the observation set and time stamps are provided.

In this paper, we focus on learning object-centric and viewpoint representations conditioned on time stamps from multi-view static scenes for video decomposition and unknown-viewpoint prediction. The problem setting and the expected outputs are illustrated in Figure 1. Under the setting that only the observation set and time stamps are provided, a generative model is developed to 1) make video decomposition based on object-centric representations; 2) reconstruct the complete shapes of partially or even completely occluded objects; and 3) predict 2D images from unknown viewpoints conditioned on known viewpoints.

To enable the abovementioned abilities, we propose a time-conditioned generative model for video decomposition and prediction. The proposed model reconstructs the complete shape of an object accurately through enhancing the disentanglement between object-centric representations and viewpoint representations, where the latent representations of time-conditioned views are jointly inferred with a Transformer [Vaswani et al., 2017] and then are input to a sequential extension of Slot Attention [Locatello et al., 2020] to learn viewpoint-invariant object-centric representations. In addition, the prediction from novel viewpoints without viewpoint annotations is enabled. Specifically, Gaussian processes are employed as priors of viewpoint latent variables for video generation and novel-view inference, based on the learned functions depicting the underlying implicit rules in view representations.

Experiments on multiple synthetic datasets demonstrate that the proposed model can 1) make object-centric video decomposition, 2) reconstruct the complete shapes of occluded objects, and 3) make novel-view predictions. Moreover, the proposed model outperforms the state-of-the-art methods in video decomposition and, compared with the method that uses viewpoint annotations, achieves competitive results on novel-view prediction.

2 RELATED WORK

Single-View Object-Centric Learning. Recent advances mainly focus on aggregating the input image into multiple slots based on the attention mechanism. AIR [Eslami et al., 2016] extracts a variable number of object representations based on the bounding-box attention [Jaderberg et al., 2015]. SQAIR [Kosiorek et al., 2018] further extends AIR to videos. Both SPACE [Lin et al., 2019] and GMIOO [Yuan et al., 2019a] model the background separately and model occlusions from different perspectives. SCALOR [Jiang et al., 2019] implements object discovery and tracking in videos with dynamic backgrounds based on SPACE. G-SWM [Lin et al., 2020] integrates the advantages of current models on videos and further models the multimodal uncertainty. MONet [Burgess et al., 2019] adopts the attention network to iteratively infer masks and then extract object-centric representations based on masked features. GENESIS [Engelcke et al., 2020] additionally models layouts of scenes based on MONet. GENESIS-V2 [Engelcke et al., 2021] infers the attention masks inspired by instance coloring previously used in supervised instance segmentation. Slot Attention [Locatello et al., 2020] and EfficientMORL [Emami et al., 2021] randomly initialize the embeddings of objects in the slots to compute the similarities between the embeddings and local features. ADI [Yuan et al., 2021] proposes a continual learning strategy and makes pilot explorations in the acquisition and exploitation of knowledge.

Multi-View Object-Centric Learning. We can coarsely categorize the recent advances in terms of viewpoint annotation. GQN [Eslami et al., 2018] uses viewpoint annotations to build single-object scenes. Based on novel-view annotations, single-object images from the given viewpoints

can be generated. MulMON [Li et al., 2020] models the multi-object multi-view scenes according to viewpoint annotations. The double-level iterative inference is conducted to achieve both multi-object segmentation and prediction. ROOTS [Chen et al., 2021] divides the three-dimensional space into equal-spaced grids and discovers objects in different grids. ROOTS also considers occlusions and makes predictions with viewpoint annotations. SIMONe [Kabra et al., 2021] and OCLOC [Yuan et al., 2022b] are the most recent models without viewpoint annotations. They learn viewpoint representations and object-centric representations separately. The difference is that SIMONe learns representations from videos and can recompose representations to novel scenes, while OCLOC is capable of modeling scenes from unordered viewpoints.

Deep Learning with Stochastic Processes. The Gaussian Process (GP) [Rasmussen and Williams, 2006] is a classical non-parametric model that regards the outputs of a function as a random variable of multivariate Gaussian distribution. The Neural Process (NP) [Garnelo et al., 2018, Kim et al., 2019] captures function stochasticity with a Gaussian distributed latent variable obtained from an inference network. To integrate stochastic processes into generative models, [Shi et al., 2021] employs GPs with deep kernels for Raven's progressive matrices completion. CLAP-NP [Shi et al., 2023] takes the first attempt in compositional law parsing with random functions based on NPs. In addition, a number of deep generative models [Deng et al., 2020, Norcliffe et al., 2021, Song et al., 2021] introduce ODEs or SDEs to learn diverse random functions on latent states.

3 BACKGROUND

In order to enable the abilities illustrated in Figure 1, in the following we list the treatments to consider in multi-view object-centric representation learning from videos without viewpoint annotations.

Variable Number of Objects. As the number of objects differs from one scene to another, it requires modeling and inference. A possible solution is to introduce a set of Bernoulli variables $\boldsymbol{z}^{\text{pres}} = \{z_1^{\text{pres}}, ..., z_K^{\text{pres}}\}$ to model object presences in the K slots for automatic counting, where K denotes the maximum number of objects that may appear in a scene.

Separately Modeling of Background. As foreground objects only occupy local regions while the background covers the entire image, the generation of 3D objects from multiple viewpoints tends to blur through a decoder shared with the background. We train two different decoders, a shared foreground object decoder and a separate background decoder.

View-independent Object Representations. We don't learn object representations from different viewpoints separately. As we can view representations of the same object inherently consistent independent of viewpoints, we consider $\{\boldsymbol{z}^{\text{bck}}, \boldsymbol{z}_1^{\text{obj}}, ..., \boldsymbol{z}_K^{\text{obj}}\}$ as view-independent object-centric representations, learned from multiple observed viewpoints to represent viewpoint-invariant 3D objects.

Depth Estimation of Objects. We introduce a depth variable $o_{t,k} \in \big[0, 1\big]$ of the kth object in the tth frame and its complete shape $\boldsymbol{s}_{t,k}^{\text{shp}} \in \big[0, 1\big]^N$ before being occluded in generative modeling. In this way, the pixels of an object with larger depth values will cover the pixels with smaller depth values. We can thus naturally obtain the observed shape of an possibly occluded object. It is worth noting that this treatment is also applicable to situations where an object is completely occluded.

Modeling of Viewpoints. We explicitly learn the viewpoint representations according to modelling the correlations of viewpoints, instead of directly leveraging viewpoint annotations as previous works [Li et al., 2020, Chen et al., 2021]. The view-correlation based modeling can also enable novel-view prediction given any time. To this end, we define $\boldsymbol{z}^{\text{view}} \in \mathbb{R}^{T\times D}$ and $\boldsymbol{\lambda} \in \mathbb{R}^{T\times D\times D_\lambda}$, where T denotes the number of frames, D denotes the dimensionality of viewpoint representations, and $\boldsymbol{z}^{\text{view}}$ follows the GPs w.r.t. $\boldsymbol{\lambda}$ that characterizes the position of the camera in different frames.

4 METHOD

Our goal is to infer object-centric latent variables independent of viewpoints and correlated viewpoint latent variables dependent on time t. In the following, we introduce our time-conditioned generative model, the inference method and a two-stage training procedure to achieve the goal.

4.1 GENERATIVE MODEL

Let $\boldsymbol{x}_{\mathcal{S}} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_T\}$ be the T frames in a video and $\boldsymbol{t}_{\mathcal{S}}$ be their timestamps. The frame set $\boldsymbol{x}_{\mathcal{S}}$ can be arbitrarily divided into an observation frame set $\boldsymbol{x}_{\mathcal{T}}$ and a prediction frame set $\boldsymbol{x}_{\mathcal{Q}}$, where $\boldsymbol{x}_{\mathcal{S}} = \boldsymbol{x}_{\mathcal{T}} \cup \boldsymbol{x}_{\mathcal{Q}}$. For convenience, the elements in $\boldsymbol{x}_{\mathcal{T}}$ and $\boldsymbol{x}_{\mathcal{Q}}$ is sorted according to the time, e.g. $\boldsymbol{x}_{\mathcal{T}} = \big(\boldsymbol{x}_1, \boldsymbol{x}_3, \boldsymbol{x}_7, \boldsymbol{x}_9\big)$; similarly, $\boldsymbol{t}_{\mathcal{S}}$ can be divided into $\boldsymbol{t}_{\mathcal{T}}$ and $\boldsymbol{t}_{\mathcal{Q}}$ accordingly. Figure 2 shows the flowchart of the generative process. The generative model conditioned on time $\boldsymbol{t}_{\mathcal{S}}$ can be expressed as:

$$\boldsymbol{\lambda}_{t,d} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{w}_t, \sigma_w^2 \boldsymbol{I}) \tag{1}$$

$$\kappa_{\boldsymbol{\eta}}^d(\boldsymbol{\lambda}_{t,d}, \boldsymbol{\lambda}_{t',d}) = l^2 \exp\Big(\frac{\|g_{\boldsymbol{\eta}}^d(\boldsymbol{\lambda}_{t,d}) - g_{\boldsymbol{\eta}}^d(\boldsymbol{\lambda}_{t',d})\|_2^2}{2\sigma^2}\Big) \tag{2}$$

$$\boldsymbol{z}_k^{\text{obj}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \qquad \boldsymbol{z}^{\text{bck}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{3}$$

$$\boldsymbol{K}_{\boldsymbol{\eta}}^d = \begin{bmatrix} \kappa_{\boldsymbol{\eta}}^d(\boldsymbol{\lambda}_{1,d}, \boldsymbol{\lambda}_{1,d}) & \cdots & \kappa_{\boldsymbol{\eta}}^d(\boldsymbol{\lambda}_{1,d}, \boldsymbol{\lambda}_{T,d}) \\ \vdots & \ddots & \vdots \\ \kappa_{\boldsymbol{\eta}}^d(\boldsymbol{\lambda}_{T,d}, \boldsymbol{\lambda}_{1,d}) & \cdots & \kappa_{\boldsymbol{\eta}}^d(\boldsymbol{\lambda}_{T,d}, \boldsymbol{\lambda}_{T,d}) \end{bmatrix} \tag{4}$$

$$\boldsymbol{z}_{1:T,d}^{\text{view}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}_{\boldsymbol{\eta}}^d) \tag{5}$$

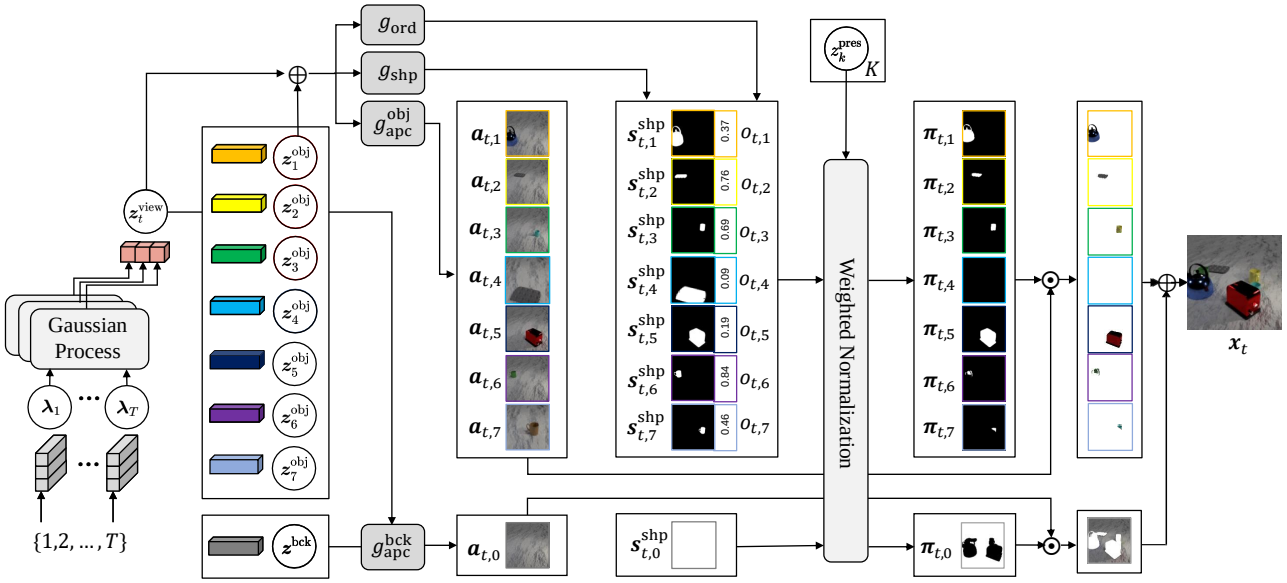

Figure 2: The proposed time-conditioned generative process for generating the tth frame in a video. The correlations between the viewpoint representations of T frames are modeled dimension-wisely with GPs. The notations in circles denote latent variables; the notations in deep gray boxes denote neural networks.

$$\boldsymbol{z}^{\text{view}}_{1:T} = \text{concat}(\boldsymbol{z}^{\text{view}}_{:,1}, ..., \boldsymbol{z}^{\text{view}}_{:,D}) \tag{6}$$

$$z^{\text{pres}}_k \sim \text{Bernoulli}(\nu_k) \qquad \nu_k \sim \text{Beta}(\alpha/K, 1) \tag{7}$$

$$s^{\text{shp}}_{t,k,n} = \text{Sigmoid}(g_{\text{shp}}(\boldsymbol{z}^{\text{obj}}_k, \boldsymbol{z}^{\text{view}}_t)_n) \tag{8}$$

$$o_{t,k} = g_{\text{ord}}(\boldsymbol{z}^{\text{obj}}_k, \boldsymbol{z}^{\text{view}}_t) \tag{9}$$

$$\pi_{t,k,n} = \begin{cases} \prod_{k'=1}^{K}(1 - z^{\text{pres}}_{k'} s^{\text{shp}}_{t,k',n}), & k = 0 \\ \frac{(1-\pi_{t,0,n})(1-z^{\text{pres}}_k s^{\text{shp}}_{t,k,n} o_{t,k})}{\sum_{k'=1}^{K}(1-z^{\text{pres}}_{k'} s^{\text{shp}}_{t,k',n} o_{t,k'})}, & k \geq 1 \end{cases} \tag{10}$$

$$\boldsymbol{a}_{t,k,n} = \begin{cases} g^{\text{bck}}_{\text{apc}}(\boldsymbol{z}^{\text{view}}_t, \boldsymbol{z}^{\text{bck}})_n, & k = 0 \\ g^{\text{obj}}_{\text{apc}}(\boldsymbol{z}^{\text{view}}_t, \boldsymbol{z}^{\text{obj}}_k)_n, & k \geq 1 \end{cases} \tag{11}$$

$$\boldsymbol{x}_{t,n} \sim \mathcal{N}\Big(\sum_{k=0}^{K} \pi_{t,k,n} \boldsymbol{a}_{t,k,n}, \sigma_x^2 \boldsymbol{I}\Big) \tag{12}$$

In the above, the ranges of all indices ($1 \leq t \leq T, 1 \leq d \leq D, 1 \leq k \leq K, 1 \leq n \leq N$) are omitted for simplicity. The way to time embedding $\boldsymbol{w}_t = \text{TimeEncoding}(t)$ can be diverse, e.g. $\boldsymbol{w}_t = \big[\cos t, \sin t\big]$. $\boldsymbol{\lambda}_{t,d}$ follows a linear Gaussian distribution with a projection matrix $\boldsymbol{A}$, which can be either learned or provided, and σ_w is a hyperparameter. $\kappa^d_{\boldsymbol{\eta}}$ is the kernel function corresponding to the dth dimension of $\boldsymbol{z}^{\text{view}}$ composed of a neural network $g^d_{\boldsymbol{\eta}}$ and an RBF kernel parameterized with $\boldsymbol{\eta}$, l and σ ([Wilson et al., 2016]). Each dimension of the viewpoint latent variable $\boldsymbol{z}^{\text{view}}_t$ is generated by a different GP in Eq.5. The occlusions are treated in Eq.10 through sorting the depth values of objects to obtain the soft masks $\boldsymbol{\pi}_{t,k}$ of the background and objects. $\boldsymbol{a}_{t,k}$ in Eq.11 denotes the complete appearance of the kth object or background in GRB values at time t. The likelihood of the nth observed pixel at time t is a Gaussian distribution parameterized with $\boldsymbol{\pi}$ and $\boldsymbol{a}$ in Eq.12.

Let $\boldsymbol{\Omega} = \{\boldsymbol{z}^{\text{obj}}, \boldsymbol{z}^{\text{bck}}, \boldsymbol{z}^{\text{pres}}, \boldsymbol{z}^{\text{view}}, \boldsymbol{\lambda}, \boldsymbol{\nu}\}$ denote the collection of all latent variables, the joint conditional probability of $\boldsymbol{x}_{\mathcal{S}}$ and $\boldsymbol{\Omega}$ can be written as:

$$\begin{aligned} p(\boldsymbol{x}_{\mathcal{S}}, \boldsymbol{\Omega} \,|\boldsymbol{t}_{\mathcal{S}}) = &\prod_{t=1}^{T}\prod_{n=1}^{N} p(\boldsymbol{x}_{t,n} \mid \boldsymbol{\Omega}) p(\boldsymbol{z}^{\text{bck}}) \\ &\cdot \prod_{d=1}^{D} p(\boldsymbol{z}^{\text{view}}_{\mathcal{S},d} \mid \boldsymbol{\lambda}_{\mathcal{S},d}) \prod_{t=1}^{T} p(\boldsymbol{\lambda}_{t,d} \mid \boldsymbol{t}_{\mathcal{S}}) \\ &\cdot \prod_{k=1}^{K} p(\boldsymbol{z}^{\text{obj}}_k) p(z^{\text{pres}}_k \mid \nu_k) p(\nu_k) \end{aligned} \tag{13}$$

4.2 INFERENCE

Since we can hardly compute the likelihood through integrating out the latent variables $\boldsymbol{\Omega}$, the amortized variational inference approach is employed to approximate the posterior of $\boldsymbol{\Omega}$. In our problem setting, only a subset of the frame collection, $\boldsymbol{x}_{\mathcal{T}}$, for each video is observed. This implies that the posteriors of $\boldsymbol{\lambda}_{\mathcal{T}}$ and $\boldsymbol{z}^{\text{view}}_{\mathcal{T}}$ that correspond to $\boldsymbol{x}_{\mathcal{T}}$ can be inferred directly with the inference networks, while the posteriors of $\boldsymbol{\lambda}_{\mathcal{Q}}$ and $\boldsymbol{z}^{\text{view}}_{\mathcal{Q}}$ that correspond to $\boldsymbol{x}_{\mathcal{Q}}$ are hard to compute. We use the least square method to approximate the posterior of $\boldsymbol{\lambda}_{\mathcal{Q}}$ and then explicitly compute the posterior of $\boldsymbol{z}^{\text{view}}_{\mathcal{Q}}$ based on the properties of the GP prior. For simplicity, the parameters in the inference networks are denoted by $\boldsymbol{\phi}$ and the parameters in the learnable kernels in GP are denoted by $\boldsymbol{\eta}$. The variational posterior $q_{\boldsymbol{\phi},\boldsymbol{\eta}}(\boldsymbol{\Omega} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})$ conditioned on the observed set can be written as:

$$\begin{aligned} q_{\boldsymbol{\phi},\boldsymbol{\eta}}(\boldsymbol{\Omega} \,|\boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}}) = &\, q_{\boldsymbol{\phi}}(\boldsymbol{z}^{\text{bck}} \mid \boldsymbol{x}_{\mathcal{T}}) q_{\boldsymbol{\phi}}(\boldsymbol{z}^{\text{view}}_{\mathcal{T}} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{T}}) \\ &\cdot q_{\boldsymbol{\phi}}(\boldsymbol{\lambda}_{\mathcal{T}} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{T}}) q_{\boldsymbol{\phi}}(\boldsymbol{\lambda}_{\mathcal{Q}} \mid \boldsymbol{\lambda}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}}) \\ &\cdot \prod_{k=1}^{K} q_{\boldsymbol{\phi}}(\boldsymbol{z}^{\text{obj}}_k \mid \boldsymbol{x}_{\mathcal{T}}) q_{\boldsymbol{\phi}}(z^{\text{pres}}_k \mid \boldsymbol{x}_{\mathcal{T}}) q_{\boldsymbol{\phi}}(\nu_k \mid \boldsymbol{x}_{\mathcal{T}}) \end{aligned}$$

$$\cdot \prod_{d=1}^{D} q_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{Q},d}^{\text{view}} \mid \boldsymbol{z}_{\mathcal{T},d}^{\text{view}}, \boldsymbol{\lambda}_{\mathcal{S},d}) \tag{14}$$

In the following, we will introduce the inference methods for the observed view-dependent latent variables in Section 4.2.1, the predicted view-dependent latent variables in Section 4.2.2, and the view-independent object-centric latent variables in Section 4.2.3. The overview of the inference procedure is illustrated in Figure 3. The mathematical details of the inference procedure can be found in the Supplementary Material.

4.2.1 Inference of Observed View-dependent Latents

The posteriors of the viewpoint latent variable $\boldsymbol{z}_t^{\text{view}}$ $(t \in \mathcal{T})$ and the timestamp latent variable $\boldsymbol{\lambda}_{t,d}$ $(t \in \mathcal{T}, 1 \leq d \leq D)$ are defined as:

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}_t^{\text{view}} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{T}}) = \mathcal{N}(\boldsymbol{z}_t^{\text{view}} \mid \boldsymbol{\mu}_t^{\text{view}}, \text{diag}(\boldsymbol{\sigma}_t^{\text{view}})^2)$$
$$q_{\boldsymbol{\phi}}(\boldsymbol{\lambda}_{t,d} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{T}}) = \mathcal{N}(\boldsymbol{\lambda}_{t,d} \mid \boldsymbol{\mu}_{t,d}^{\lambda}, \sigma_{\boldsymbol{w}}^2 \boldsymbol{I})$$

where $[\boldsymbol{\mu}_t^{\text{view}}, \boldsymbol{\sigma}_t^{\text{view}}] = f_{\boldsymbol{\phi}}^{\text{view}}(\boldsymbol{x}_{\mathcal{T}})$ and $\boldsymbol{\mu}_{t,d}^{\lambda} = f_{\boldsymbol{\phi}}^{\lambda}(\boldsymbol{x}_{\mathcal{T}}, \boldsymbol{w}_t)$; the variance $\boldsymbol{\sigma}_{\boldsymbol{w}}^2$ is fixed. As Figure 3 shows: First, $\boldsymbol{x}_{\mathcal{T}}$ is fed into a Transformer block along with a 3D position embedding [Kabra et al., 2021], where the viewpoint information with correlations between frames is learned. A $|\mathcal{T}| \times L \times C$ feature map extracted by the Transformer is averaged over $L = HW$ pixels on the feature map to obtain $\boldsymbol{y}_t^{\text{view}}$ $(t \in \mathcal{T})$, and $\boldsymbol{y}_t^{\text{view}}$ is an intermediate variable to obtain $[\boldsymbol{\mu}_t^{\text{view}}, \boldsymbol{\sigma}_t^{\text{view}}]$ and $\boldsymbol{\mu}_{t,d}^{\lambda}$ in $f_{\boldsymbol{\phi}}^{\text{view}}$ and $f_{\boldsymbol{\phi}}^{\lambda}$, respectively.

4.2.2 Inference of Predicted View-dependent Latents

Inference of latent variables related to predicted viewpoints is challenging because $\boldsymbol{x}_{\mathcal{Q}}$ is not provided. Therefore, the predicted view-dependent latent variables need to be inferred through the observed viewpoints. We introduce the inference methods for $\boldsymbol{\lambda}_{\mathcal{Q}}$ and $\boldsymbol{z}_{\mathcal{Q}}^{\text{view}}$, respectively.

Inference of $\boldsymbol{\lambda}_{\mathcal{Q}}$. According to the prior distribution of $\boldsymbol{\lambda}_{t,d}$ defined in Eq.1, $\boldsymbol{\mu}_{t,d}^{\lambda}$ of the posterior $q_{\boldsymbol{\phi}}(\boldsymbol{\lambda}_{t,d} \mid \boldsymbol{\lambda}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{T}})$ can be approximated to satisfy a linear function w.r.t. $\boldsymbol{w}_t$, i.e. $\boldsymbol{\mu}_{t,d}^{\lambda} = \hat{\boldsymbol{A}}_d \boldsymbol{w}_t, \hat{\boldsymbol{A}}_d \in \mathbb{R}^{D_\lambda \times |\boldsymbol{w}_t|}$. Based on the Least Square method, the optimal $\hat{\boldsymbol{A}}_d^*$ $(1 \leq d \leq D)$ in the linear set and the posterior of $\boldsymbol{\lambda}_{t,d}$ $(t \in \mathcal{Q})$ are:

$$q_{\boldsymbol{\phi}}(\boldsymbol{\lambda}_{t,d} \mid \boldsymbol{\lambda}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}}) = \mathcal{N}(\hat{\boldsymbol{A}}_d^* \boldsymbol{w}_t, \sigma_{\boldsymbol{w}}^2 \boldsymbol{I}) \tag{15}$$
$$\hat{\boldsymbol{A}}_d^* = \boldsymbol{\Phi}_d^\top \boldsymbol{W}_{\mathcal{T}} (\boldsymbol{W}_{\mathcal{T}}^\top \boldsymbol{W}_{\mathcal{T}})^{-1} \tag{16}$$

where $\boldsymbol{W}_{\mathcal{T}} = \left[\boldsymbol{w}_1, ..., \boldsymbol{w}_{|\mathcal{T}|}\right]^\top \in \mathbb{R}^{|\mathcal{T}| \times |\boldsymbol{w}_t|}$ and $\boldsymbol{\Phi}_d = \left[\boldsymbol{\mu}_{1,d}, ..., \boldsymbol{\mu}_{|\mathcal{T}|,d}\right]^\top \in \mathbb{R}^{|\mathcal{T}| \times D_\lambda}$.

Inference of $\boldsymbol{z}_{\mathcal{Q}}^{\text{view}}$. $q_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{Q}}^{\text{view}} \mid \boldsymbol{z}_{\mathcal{T}}^{\text{view}}, \boldsymbol{\lambda}_{\mathcal{S}})$ follows the same distribution as the predictive distribution of the GPs (the details can be found in the Supplementary Material):

$$q_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{Q}}^{\text{view}} \mid \boldsymbol{z}_{\mathcal{T}}^{\text{view}}, \boldsymbol{\lambda}_{\mathcal{S}}) = \prod_{d=1}^{D} p_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{Q},d}^{\text{view}} \mid \boldsymbol{z}_{\mathcal{T},d}^{\text{view}}, \boldsymbol{\lambda}_{\mathcal{S},d}) \tag{17}$$

where $p_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{Q},d}^{\text{view}} \mid \cdot)$ satisfies the multivariate Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_{\mathcal{Q},d}^{\text{view}}, \boldsymbol{\Sigma}_{\mathcal{Q},d}^{\text{view}})$, and the parameters $\boldsymbol{\mu}_{\mathcal{Q},d}^{\text{view}}$ and $\boldsymbol{\Sigma}_{\mathcal{Q},d}^{\text{view}}$ are analytical functions of $\boldsymbol{\lambda}_{\mathcal{S},d}$, $\boldsymbol{z}_{\mathcal{T},d}^{\text{view}}$ and $\boldsymbol{\eta}$.

4.2.3 Inference of View-independent Latents

The posteriors of the view-independent object-centric latent variables $\{\boldsymbol{z}^{\text{bck}}, \boldsymbol{z}^{\text{obj}}, \boldsymbol{z}^{\text{pres}}, \boldsymbol{\nu}\}$ in Eq.14 are defined as:

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}^{\text{bck}} \mid \boldsymbol{x}_{\mathcal{T}}) = \mathcal{N}(\boldsymbol{z}^{\text{bck}} \mid \boldsymbol{\mu}^{\text{bck}}, \text{diag}(\boldsymbol{\sigma}^{\text{bck}})^2) \tag{18}$$
$$q_{\boldsymbol{\phi}}(\boldsymbol{z}_k^{\text{obj}} \mid \boldsymbol{x}_{\mathcal{T}}) = \mathcal{N}(\boldsymbol{z}_k^{\text{obj}} \mid \boldsymbol{\mu}_k^{\text{obj}}, \text{diag}(\boldsymbol{\sigma}_k^{\text{obj}})^2) \tag{19}$$
$$q_{\boldsymbol{\phi}}(z_k^{\text{pres}} \mid \boldsymbol{x}_{\mathcal{T}}) = \text{Bernoulli}(z_k^{\text{pres}} \mid \kappa_k) \tag{20}$$
$$q_{\boldsymbol{\phi}}(\nu_k \mid \boldsymbol{x}_{\mathcal{T}}) = \text{Beta}(\nu_k \mid \tau_{k,1}, \tau_{k,2}) \tag{21}$$

where the default range of k is $1 \leq k \leq K$. All the parameters of the above distributions will pass through a sequential extension of Slot Attention [Locatello et al., 2020], which is illustrated in Figure 3.

The model maintains $K+1$ slots $\boldsymbol{y}^{\text{attr}} = [\boldsymbol{y}^{\text{bck}}, \boldsymbol{y}_1^{\text{obj}}, ..., \boldsymbol{y}_K^{\text{obj}}]$, $\boldsymbol{y}_k^{\text{attr}} \in \mathbb{R}^{D_s}$. Different from Slot Attention [Locatello et al., 2020], two types of initialization are employed for the foreground objects and the background, respectively. Then $\boldsymbol{y}_k^{\text{attr}}$ is combined with $\boldsymbol{y}_t^{\text{view}} \in \mathbb{R}^{D_v}$ $(t \in \mathcal{T})$ obtained in Section 4.2.1 to produce $|\mathcal{T}| \times (K+1)$ slots $\boldsymbol{y}_{t,k}^{\text{full}} \in \mathbb{R}^{D_f}$ with the viewpoint information, where $D_f = D_s + D_v$. We use another encoder to extract the feature maps of $\boldsymbol{x}_{\mathcal{T}}$, denoted as $\boldsymbol{y}_{\mathcal{T}}^{\text{sa}}$. We do M iterations like Slot Attention. In each iteration, Eq.22 first uses the cross attention to obtain the attention masks $\boldsymbol{a}_t \in \mathbb{R}^{N \times (K+1)}$ of K objects and the background. Then, the pixel-wise normalized masks of all the objects and background are multiplied with the value of $\boldsymbol{y}_t^{\text{sa}}$ to obtain the hidden state $\boldsymbol{u}_t \in \mathbb{R}^{(K+1) \times D_f}$ for GRU updating. In addition, we perform temporal mean over the updated attribute part of $\hat{\boldsymbol{y}}_{t,k}^{\text{full}}$ after GRU updating.

$$\boldsymbol{a}_t = \underset{K+1}{\text{Softmax}}\Big(\frac{k(\boldsymbol{y}_t^{\text{sa}}) \cdot q(\boldsymbol{y}_{t,1:K+1}^{\text{full}})^\top}{\sqrt{D_f}}\Big) \tag{22}$$

$$\boldsymbol{u}_t = \sum_{n=1}^{N} \Big(\underset{N}{\text{Softmax}}\big(\log \boldsymbol{a}_{t,n}\big) \cdot v(\boldsymbol{y}_{t,n}^{\text{sa}})\Big) \tag{23}$$

$$\hat{\boldsymbol{y}}_{t,k}^{\text{full}} = \text{GRU}(\boldsymbol{y}_{t,k}^{\text{full}}, \boldsymbol{u}_{t,k}) \quad \left[\hat{\boldsymbol{y}}_{t,k}^{\text{attr}}, \hat{\boldsymbol{y}}_{t,k}^{\text{view}}\right] \overset{\text{split}}{\leftarrow} \hat{\boldsymbol{y}}_{t,k}^{\text{full}} \tag{24}$$

$$\boldsymbol{y}_k^{\text{attr}} = \text{mean}_{|\mathcal{T}|}\Big(\hat{\boldsymbol{y}}_{1:|\mathcal{T}|,k}^{\text{attr}}\Big) \tag{25}$$

where k, q and v are MLPs for producing key, query and value, respectively. The procedure maintains the permutation invariance w.r.t. the input order of frames. $\boldsymbol{\mu}^{\text{bck}}$ and $\boldsymbol{\sigma}^{\text{bck}}$ are obtained through the neural network $f_{\boldsymbol{\phi}}^{\text{bck}}$ with $\boldsymbol{y}^{\text{bck}}$ as input; $\boldsymbol{\mu}_k^{\text{obj}}, \boldsymbol{\sigma}_k^{\text{obj}}, \kappa_k, \tau_{k,1}, \tau_{k,2}$ are obtained through the shared neural network $f_{\boldsymbol{\phi}}^{\text{obj}}$ with $\boldsymbol{y}_k^{\text{obj}}$ as input.

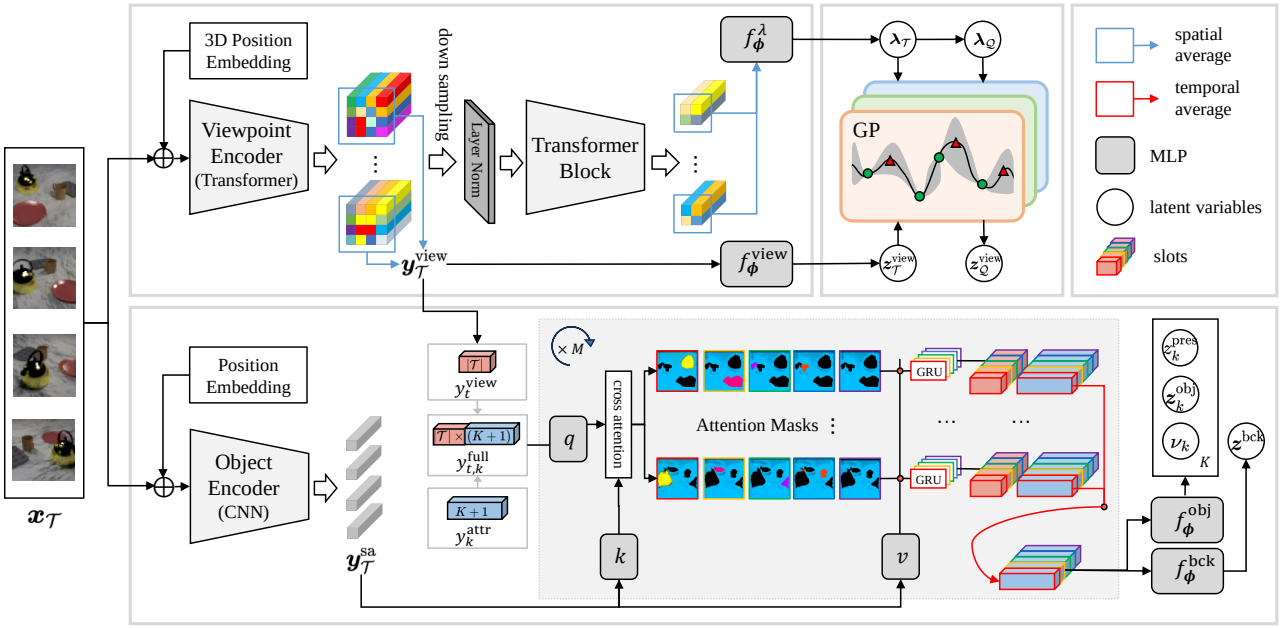

Figure 3: The inference procedure of the proposed model. The three modules correspond to the inference of observed view-dependent latent variables (top-left), the inference of predicted view-dependent latent variables (top-middle), and the inference of view-independent object-centric latent variables (bottom), respectively.

4.3 TRAINING

Optimizing the evidence lower bound (ELBO) for all frames (including both observed and predicted frames) is unstable. To solve this problem, a two-stage training procedure is adopted. Let $\boldsymbol{\Omega}_{\mathcal{S}} = \{\boldsymbol{\Omega}_{\mathcal{T}}, \boldsymbol{\Omega}_{\mathcal{Q}}\}$, where $\boldsymbol{\Omega}_{\mathcal{T}} = \{\boldsymbol{z}^{\text{bck}}, \boldsymbol{z}^{\text{obj}}, \boldsymbol{z}^{\text{pres}}, \boldsymbol{\nu}, \boldsymbol{\lambda}_{\mathcal{T}}, \boldsymbol{z}^{\text{view}}_{\mathcal{T}}\}$ and $\boldsymbol{\Omega}_{\mathcal{Q}} = \{\boldsymbol{z}^{\text{bck}}, \boldsymbol{z}^{\text{obj}}, \boldsymbol{z}^{\text{pres}}, \boldsymbol{\nu}, \boldsymbol{\lambda}_{\mathcal{Q}}, \boldsymbol{z}^{\text{view}}_{\mathcal{Q}}\}$, i.e. the view-independent latent variables share in both $\boldsymbol{\Omega}_{\mathcal{T}}$ and $\boldsymbol{\Omega}_{\mathcal{Q}}$. The two-stage losses are as follows:

$$\begin{aligned}\mathcal{L}_1 = &- \mathbb{E}_{q_{\boldsymbol{\phi},\boldsymbol{\eta}}(\boldsymbol{\Omega}_{\mathcal{T}}|\boldsymbol{x}_{\mathcal{T}})}\left[\log p_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{x}_{\mathcal{T}} \mid \boldsymbol{\Omega}_{\mathcal{T}})\right] \\ &+ D_{KL}\Big(q_{\boldsymbol{\phi},\boldsymbol{\eta}}(\boldsymbol{\Omega}_{\mathcal{T}} \mid \boldsymbol{x}_{\mathcal{T}}) \| p_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{\Omega}_{\mathcal{T}})\Big) \qquad (26)\end{aligned}$$

$$\begin{aligned}\mathcal{L}_2 = &- \frac{1}{|\mathcal{T}|}\mathbb{E}_{q_{\boldsymbol{\phi},\boldsymbol{\eta}}(\boldsymbol{\Omega}_{\mathcal{T}}|\boldsymbol{x}_{\mathcal{T}},\boldsymbol{t}_{\mathcal{T}})}\left[\log p_{\boldsymbol{\phi},\boldsymbol{\eta}}(\boldsymbol{x}_{\mathcal{T}} \mid \boldsymbol{\Omega}_{\mathcal{T}})\right] \\ &- \frac{1}{|\mathcal{Q}|}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\Omega}_{\mathcal{T}}|\boldsymbol{x}_{\mathcal{T}},\boldsymbol{t}_{\mathcal{T}})q_{\boldsymbol{\phi},\boldsymbol{\eta}}(\boldsymbol{\Omega}_{\mathcal{Q}}|\boldsymbol{\Omega}_{\mathcal{T}},\boldsymbol{t}_{\mathcal{Q}})}\left[\log p_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{x}_{\mathcal{Q}} \mid \boldsymbol{\Omega}_{\mathcal{Q}})\right] \\ &+ \beta D_{KL}\Big(q_{\boldsymbol{\phi},\boldsymbol{\eta}}(\boldsymbol{\Omega}_{\mathcal{S}} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}}) \| p_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{\Omega}_{\mathcal{S}} \mid \boldsymbol{t}_{\mathcal{S}})\Big) \qquad (27)\end{aligned}$$

where $\mathcal{L}_1$ is a standard ELBO of $\boldsymbol{\Omega}_{\mathcal{T}}$ on $\boldsymbol{x}_{\mathcal{T}}$ to learn object-centric representations from multiple frames and does not depend on $\boldsymbol{t}_{\mathcal{S}}$; while $\mathcal{L}_2$ adopts the curriculum learning to learn the function of viewpoint latent variables w.r.t. $\boldsymbol{t}_{\mathcal{S}}$. Let $\mathcal{S}'$ denote the subset of $\mathcal{S}$ and $|\mathcal{S}'|$ is scheduled to gradually increase during training. $\mathcal{S}'$ will be randomly divided into $\mathcal{T}$ and $\mathcal{Q}$, where $|\mathcal{Q}| \sim U(1, C)$ ($C < |\mathcal{S}'|$ and increases during training). $\mathcal{L}_2$ averages the observed and predicted losses to balance the two losses, where $\beta \geq 1$ is a hyper-parameter follows [Burgess et al., 2018]. Note that the reconstruction performance of $\mathcal{L}_2$ is worse than that of the first stage; however, it can perform well on the prediction task.

5 EXPERIMENTS

We design experiments to investigate 1) how well the proposed model performs compared to state-of-the-art multi-view models in object-centric video decomposition on the observation set; 2) whether the proposed model can disentangle the 3D scene into object-centric view-invariant representations and viewpoint representations; 3) how well the proposed model handles occlusions compared to existing methods; 4) how well the proposed model makes the prediction only depending on timestamps; and 5) whether the proposed model can generate videos.

To validate the above, we compare the proposed model[1] with three state-of-the-art models, **MulMON** [Li et al., 2020] with viewpoint annotations, viewpoint-free models **SIMONe** [Kabra et al., 2021] and **OCLOC** [Yuan et al., 2022b]. We design four synthetic video datasets, called CLEVR-SIMPLE, CLEVR-COMPLE, SHOP-SIMPLEX, and SHOP-COMPLEX, through modifying multi-view CLEVR [Johnson et al., 2017] and SHOP [Nazarczuk and Mikolajczyk, 2020] based on the official code. The two SHOP datasets are more challenging than the two CLEVR

[1]The code is available at https://github.com/FudanVI/compositional-scene-representation-toolbox

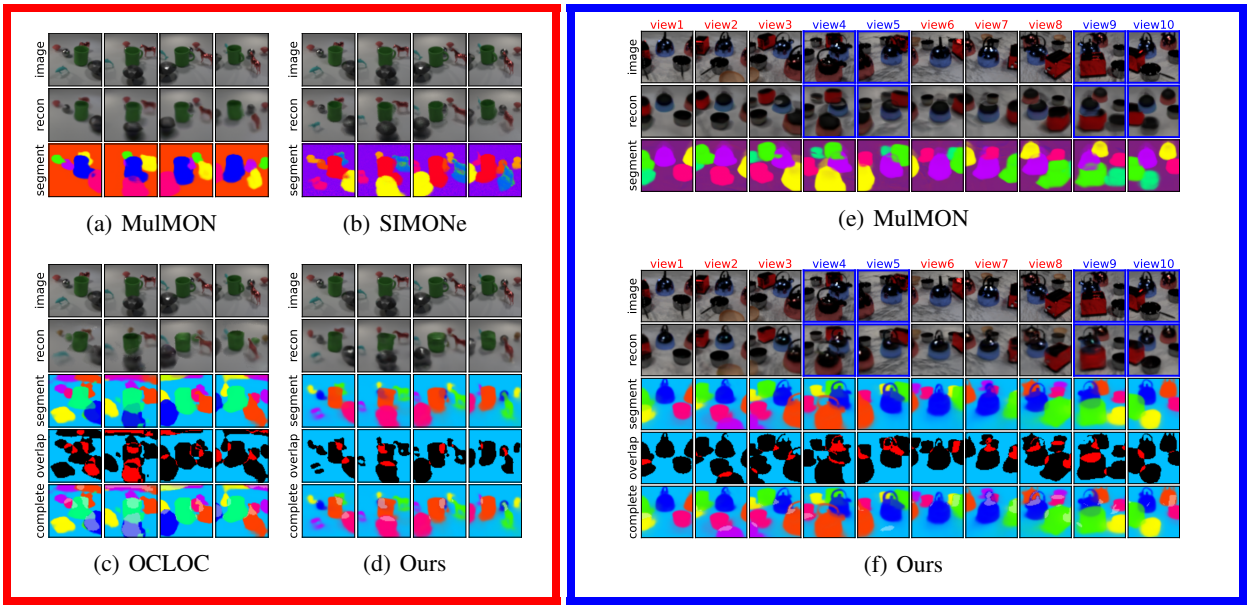

Figure 4: **Left**: Visualization results of the compared methods on the *observation* set of CLEVER-COMPLEX, where four consecutive frames are demonstrated. **Right**: Visualization results on the *prediction* set of SHOP-SIMPLE. The 'images' in blue boxes are unobserved ground truths and the 'recons' in blue boxes are predicted results.

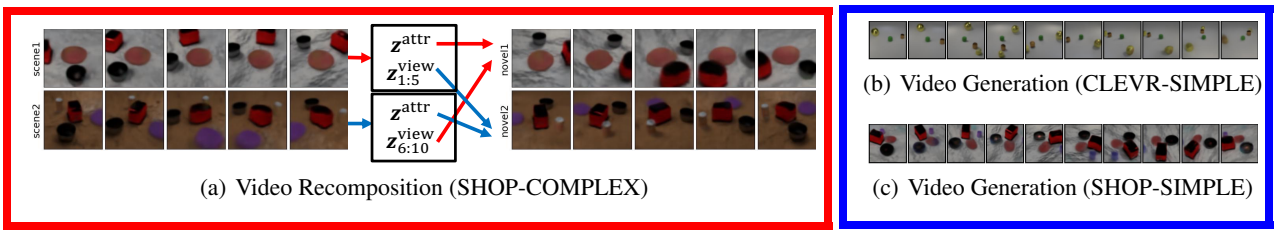

Figure 5: **Left**: Scene image generation from novel viewpoints through recomposing viewpoint representations and object-centric representations. **Right**: Video generation based on CLEVR-SIMPLE and SHOP-SIMPLE.

datasets in terms of the object texture; the two COMPLEX versions are more challenging than the two SIMPLE versions because of more types of objects and backgrounds.

We train the proposed model with the introduced two-stage strategy. Stage 1 can reconstruct the observation set without supervision while Stage 2 can predict unobserved set only with timestamp supervision. We train the proposed model on all the datasets using the Adam optimizer with a learning rate 4e-4 for 300K gradient steps. The increment of curriculum learning is 2.

Video Decomposition. Since the proposed model maintains the view-invariant object-centric representations in 3D structure, video decomposition is crucial to evaluating the completeness and accuracy of learned representations. Figure 4 (Left) demonstrates the visualization results on CLEVR-COMPLEX. The proposed model can accurately represent objects with complex shapes from multiple viewpoints and build crisp segregation between the foregrounds and the background. Moreover, the proposed model tends to treat shadows as parts of objects (e.g., the horse in Figure 4(d)), it is reasonable for shadows to be blended with the corresponding objects due to lighting. Surprisingly, the shadow area is noticeably smaller than those of other models.

Table 1(a) reports the segmentation performance in terms of foreground objects. ARI-O measures how accurately a video is decomposed into separate objects. We find that, except for CLEVR-SIMPLE, the proposed model outperforms the other models, especially on the two SHOP datasets, probably because the 3D representations integrity of objects helps reconstruct better masks. SIMONe and OCLOC fail to capture the objects on SHOP-COMPLEX. A possible reason is that the background is indistinguishable with the objects in SHOP-COMPLEX, such that these models cannot represent the background separately during the inference. Although OCLOC models the background separately, sampling from permutation-equivalent slots may affect the extraction of the background representation.

Video Recomposition. An intriguing experiment is to gen-

Table 1: Performance comparison of MulMON, SIMONe and the proposed model (Ours). ARI-O is adopted for evaluating segmentation, IoU and OOA are adopted for evaluating segmentation with occlusions, and MSE is adopted for evaluating reconstruction. Except for MSE in (d), all results are recorded in 'mean ± std' over 5 random seeds. '-S' and '-C' are short for 'SIMPLE' and 'COMPLEX', respectively.

(a) ARI-O (observation set)

Model	CLEVR-S	CLEVR-C	SHOP-S	SHOP-C
	ARI-O↑	ARI-O↑	ARI-O↑	ARI-O↑
MulMON (cond)	**96.4** ± 0.1	92.9 ± 0.2	88.3 ± 0.6	87.1 ± 0.2
SIMONe	91.0 ± 0.0	91.4 ± 0.0	55.3 ± 0.0	33.5 ± 0.0
OCLOC	92.7 ± 0.8	82.7 ± 0.8	91.3 ± 0.4	29.3 ± 0.5
Ours	95.9 ± 0.3	**94.1** ± 0.3	**95.8** ± 0.1	**94.9** ± 0.4

(b) IoU and OOA (observation set)

Model	IoU↑		OOA↑	
	OCLOC	Ours	OCLOC	Ours
CLEVR-S	45.6 ± 0.2	**59.5** ± 0.5	93.6 ± 1.2	**95.3** ± 1.1
CLEVR-C	35.1 ± 0.2	**50.9** ± 0.4	89.1 ± 1.2	**93.0** ± 0.8
SHOP-S	61.9 ± 0.6	**65.9** ± 0.1	72.8 ± 1.4	**78.9** ± 0.4
SHOP-C	21.5 ± 0.3	**66.2** ± 0.6	57.9 ± 1.9	**81.8** ± 1.3

(c) ARI-O (prediction set)

Model		CLEVR-S	CLEVR-C	SHOP-S	SHOP-C
		ARI-O↑	ARI-O↑	ARI-O↑	ARI-O↑
Mode 1	MulMON	**96.2** ± 0.1	91.5 ± 0.3	88.3 ± 0.5	86.9 ± 0.7
	Ours	95.5 ± 0.5	**95.5** ± 0.9	**96.0** ± 0.3	**92.9** ± 0.4
Mode 2	MulMON	**96.9** ± 0.2	94.5 ± 0.2	87.1 ± 0.6	86.0 ± 0.6
	Ours	95.1 ± 0.5	**95.0** ± 0.6	**95.5** ± 0.1	**93.8** ± 0.8

(d) MSE (prediction set)

Model		CLEVR-S	CLEVR-C	SHOP-S	SHOP-C
		MSE↓	MSE↓	MSE↓	MSE↓
Mode 1	MulMON	**0.0014**	**0.0020**	0.0049	0.0038
	Ours	0.0018	0.0021	**0.0034**	**0.0036**
Mode 2	MulMON	**0.0014**	**0.0020**	0.0050	0.0038
	Ours	0.0017	0.0024	**0.0035**	**0.0038**

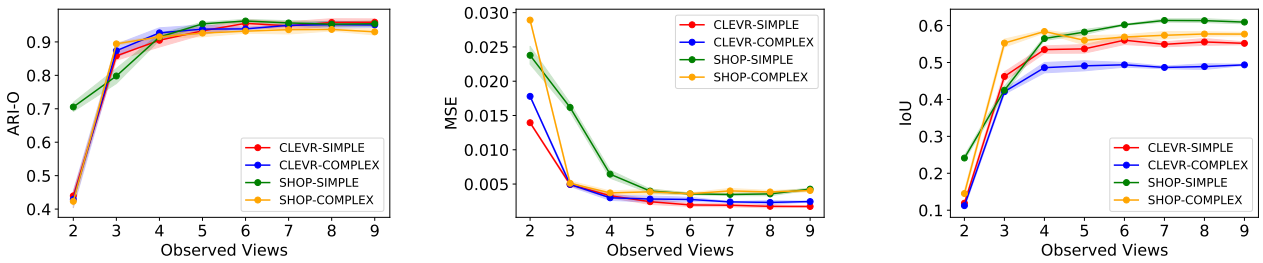

Figure 6: Single-view prediction performance in ARI-O, MSE, and IoU in terms of the number of observed views. All results are tested with 5 random seeds and each point on a curve is the mean value and the shaded band denotes ±std.

erate scene images from novel viewpoints through cross-combining viewpoint representations and object-centric representations of objects (including z^{bck} and z^{obj}). The recomposition is implemented as follows: We randomly choose two videos (each comprises 10 frames) and select the first 5 frames from one video and select the last 5 frames from the other. Then, we encode the selected frames into viewpoint representations and object-centric representations. Finally, we combine the first five object-centric representations from one video and the last five viewpoint representations from the other frame-wisely to generate the scene images from novel viewpoints. Figure 5(a) demonstrates that disentangled object-centric and viewpoint representations from different scenes can be effectively coupled, based on which the proposed model can generate novel views.

Occlusion Evaluation. Among the compared methods, only OCLOC is designed to handle occlusions. The comparison results on CLEVR-COMPLEX are visualized in Figure 4 (c) and (d). As the camera moves counterclockwise around the center, a gray ball is completely occluded behind the green mug in the second frame. The proposed model can reconstruct the complete shape of an object even it is completely occluded (e.g. the gray ball). We evaluate IoU and OOA used in [Yuan et al., 2019a] that respectively assess the quality of reconstructed complete shapes and the accuracy of the estimated pairwise ordering of objects. The proposed model clearly outperforms OCLOC, probably because OCLOC samples the pixel-wise shape during the generation, which produces noisy pixels and large shadows.

GP Prediction. Due to modeling the viewpoint latent variables with GPs, we can use the analytical posterior of $z_{\mathcal{Q}}^{\text{view}}$ to predict the rest viewpoints given the observation set. In our experimental setting, 10 consecutive viewpoint representations in Figure 4 satisfy the GPs and we randomly remove four frames (i.e. the ground truths in the blue boxes are unobserved). The remaining six frames are encoded to infer z^{obj}, z^{bck}, $\boldsymbol{\lambda}_{\mathcal{T}}$, $\boldsymbol{\lambda}_{\mathcal{Q}}$, $z_{\mathcal{T}}^{\text{view}}$ and $z_{\mathcal{Q}}^{\text{view}}$. The four viewpoint representations predicted by GPs are concatenated

with the object-centric representations to reconstruct the scene images. Figure 4(f) shows that the proposed model can predict arbitrary-time frames given the observation. Compared with MulMON which uses viewpoint annotations, the proposed model can additionally process occlusions while reconstructing frames from novel viewpoints. To assess the segmentation performance and reconstruction quality on the prediction set, we choose four fixed frames in Mode 1 and Mode 2 to make prediction (see the Supplementary Material for details). Table 1(c) and (d) show that the proposed model is comparable to MulMON on the two CLEVR datasets and clearly outperforms MulMON on the two SHOP datasets. The reconstruction loss helps improve the texture characterization of objects, which may be the reason that the proposed model achieves better performance in MSE on the two SHOP datasets.

Ablation Study. GPs have a generic nature: As the number of observed variables increases, the prediction uncertainty gradually decreases. We assume the number of observed frames (hyperparameter) to be the most important factor that affects the accuracy and uncertainty of the prediction. To verify the assumption, we fix a single frame and gradually increase the number of observed frames from 2 to 9. The viewpoint representations of both the predicted frame and the observed frames are used to construct GPs together. We execute the GP prediction and plot the performance curves in ARI-O, MSE, and IoU in terms of the number of observed views in Figure 6. One can see that the proposed model gradually reduces the uncertainty and improves the performance as the number of observed views increases, and tends to be stable after the number of observed views achieves 5.

Video Generation. As we model the viewpoint latent variables with GPs, we can generate videos from the GPs along the timeline. Figure 5(b) and (c) plot two example videos with 10 frames generated based on CLEVR-SIMPLE and SHOP-SIMPLE. One can find that the 10 frames obviously rotate clockwise around the center, reflecting the captured correlations between viewpoints; meanwhile, the generated objects and backgrounds have no irregular shapes.

6 CONCLUSION

We propose a time-conditioned generative model for video decomposition and prediction. The proposed model enhances the disentanglement between viewpoint and object-centric representations, and additionally adopts GPs for viewpoint modeling, inference and generation. We design experiments to show that the proposed model can: 1) aggregate 3D object-centric information from multiple viewpoints, and as a result, outperforms the state-of-art multi-view models; 2) restore the complete shapes of objects even when completely occluded; and 3) predict the scene images from unknown viewpoints without viewpoint annotations.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (No.62176060), STCSM project (No.20511100400), and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

References

Roger E Beaty, Mathias Benedek, Paul J Silvia, and Daniel L Schacter. Creative cognition and brain network dynamics. *Trends in Cognitive Sciences*, 20(2):87–95, 2016.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv:1804.03599*, 2018.

Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. MONet: Unsupervised scene decomposition and representation. *arXiv:1901.11390*, 2019.

Chang Chen, Fei Deng, and Sungjin Ahn. ROOTS: Object-centric representation and rendering of 3D scenes. *Journal of Machine Learning Research*, 22(1):11770–11805, 2021.

Ruizhi Deng, Bo Chang, Marcus A Brubaker, Greg Mori, and Andreas Lehrmann. Modeling continuous stochastic processes with dynamic normalizing flows. *Advances in Neural Information Processing Systems*, 33:7805–7815, 2020.

Patrick Emami, Pan He, Sanjay Ranka, and Anand Rangarajan. Efficient iterative amortized inference for learning symmetric and disentangled multi-object representations. In *International Conference on Machine Learning*, pages 2970–2981. PMLR, 2021.

Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner. GENESIS: Generative scene inference and sampling of object-centric latent representations. In *International Conference on Learning Representations*, 2020.

Martin Engelcke, Oiwi Parker Jones, and Ingmar Posner. GENESIS-v2: Inferring unordered object representations without iterative refinement. *Advances in Neural Information Processing Systems*, 34:8085–8094, 2021.

SM Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E. Hinton. Attend, infer, repeat: Fast scene understanding with generative models. *Advances in Neural Information Processing Systems*, 29, 2016.

SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv:1807.01622*, 2018.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial Transformer networks. *Advances in Neural Information Processing Systems*, 28, 2015.

Jindong Jiang, Sepehr Janghorbani, Gerard De Melo, and Sungjin Ahn. SCALOR: Generative world models with scalable object representations. In *International Conference on Learning Representations*, 2019.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

Scott P Johnson. How infants learn about the visual world. *Cognitive Science*, 34(7):1158–1184, 2010.

Rishabh Kabra, Daniel Zoran, Goker Erdogan, Loic Matthey, Antonia Creswell, Matt Botvinick, Alexander Lerchner, and Chris Burgess. SIMONe: View-invariant, temporally-abstracted object representations via unsupervised video decomposition. *Advances in Neural Information Processing Systems*, 34:20146–20159, 2021.

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.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. *arXiv:1901.05761*, 2019.

Adam Kosiorek, Hyunjik Kim, Yee Whye Teh, and Ingmar Posner. Sequential attend, infer, repeat: Generative modelling of moving objects. *Advances in Neural Information Processing Systems*, 31, 2018.

Nanbo Li, Cian Eastwood, and Robert Fisher. Learning object-centric representations of multi-object scenes from multiple views. *Advances in Neural Information Processing Systems*, 33:5656–5666, 2020.

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. In *International Conference on Learning Representations*, 2019.

Zhixuan Lin, Yi-Fu Wu, Skand Peri, Bofeng Fu, Jindong Jiang, and Sungjin Ahn. Improving generative imagination in object-centric world models. In *International Conference on Machine Learning*, pages 6140–6149. PMLR, 2020.

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.

Michal Nazarczuk and Krystian Mikolajczyk. SHOP-VRB: A visual reasoning benchmark for object perception. In *IEEE International Conference on Robotics and Automation*, pages 6898–6904. IEEE, 2020.

Alexander Norcliffe, Cristian Bodnar, Ben Day, Jacob Moss, and Pietro Liò. Neural ODE processes. *arXiv:2103.12413*, 2021.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*, volume 1. Springer, 2006.

Daniel L Schacter, Donna Rose Addis, Demis Hassabis, Victoria C Martin, R Nathan Spreng, and Karl K Szpunar. The future of memory: remembering, imagining, and the brain. *Neuron*, 76(4):677–694, 2012.

Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971.

Fan Shi, Bin Li, and Xiangyang Xue. Raven's progressive matrices completion with latent gaussian process priors. In *AAAI Conference on Artificial Intelligence*, pages 9612–9620, 2021.

Fan Shi, Bin Li, and Xiangyang Xue. Compositional law parsing with latent random functions. In *International Conference on Learning Representations*, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental Science*, 10(1):89–96, 2007.

Oliver H Turnbull, David P Carey, and Rosaleen A McCarthy. The neuropsychology of object constancy. *Journal of the International Neuropsychological Society*, 3(3): 288–298, 1997.

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.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial Intelligence and Statistics*, pages 370–378. PMLR, 2016.

Jinyang Yuan, Bin Li, and Xiangyang Xue. Generative modeling of infinite occluded objects for compositional scene representation. In *International Conference on Machine Learning*, pages 7222–7231. PMLR, 2019a.

Jinyang Yuan, Bin Li, and Xiangyang Xue. Spatial mixture models with learnable deep priors for perceptual grouping. In *AAAI Conference on Artificial Intelligence*, pages 9135–9142, 2019b.

Jinyang Yuan, Bin Li, and Xiangyang Xue. Knowledge-guided object discovery with acquired deep impressions. In *AAAI Conference on Artificial Intelligence*, pages 10798–10806, 2021.

Jinyang Yuan, Tonglin Chen, Bin Li, and Xiangyang Xue. Compositional scene representation learning via reconstruction: A survey. *arXiv:2202.07135*, 2022a.

Jinyang Yuan, Bin Li, and Xiangyang Xue. Unsupervised learning of compositional scene representations from multiple unspecified viewpoints. In *AAAI Conference on Artificial Intelligence*, pages 8971–8979, 2022b.