Time-Conditioned Generative Modeling of Object-Centric Representations for Video Decomposition and Prediction (Supplementary Material)

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

[1]School of Computer Science, Fudan University

1 DETAILS OF TRAINING

1.1 DERIVATION OF ELBO

$$
\begin{aligned}
\log p_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{x}_{\mathcal{S}} \mid \boldsymbol{t}_{\mathcal{S}}) &\geq \log p_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{x}_{\mathcal{S}} \mid \boldsymbol{t}_{\mathcal{S}}) - D_{KL}\big(q_{\boldsymbol{\phi},\boldsymbol{\eta}}(\boldsymbol{\Omega} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}}) \| p_{\boldsymbol{\theta}}(\boldsymbol{\Omega} \mid \boldsymbol{x}_{\mathcal{S}}, \boldsymbol{t}_{\mathcal{S}})\big) && (1)\\
&= \mathbb{E}_{q_{\boldsymbol{\phi},\boldsymbol{\eta}}(\boldsymbol{\Omega}|\boldsymbol{x}_{\mathcal{T}},\boldsymbol{t}_{\mathcal{S}})}\big[\log p_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{x}_{\mathcal{S}} \mid \boldsymbol{t}_{\mathcal{S}})\big] - \mathbb{E}_{q_{\boldsymbol{\phi},\boldsymbol{\eta}}(\boldsymbol{\Omega}|\boldsymbol{x}_{\mathcal{T}},\boldsymbol{t}_{\mathcal{S}})}\big[\log \frac{q_{\boldsymbol{\phi},\boldsymbol{\eta}}(\boldsymbol{\Omega} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})}{p_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{\Omega} \mid \boldsymbol{x}_{\mathcal{S}}, \boldsymbol{t}_{\mathcal{S}})}\big] && (2)\\
&= \mathbb{E}_{q_{\boldsymbol{\phi},\boldsymbol{\eta}}(\boldsymbol{\Omega}|\boldsymbol{x}_{\mathcal{T}},\boldsymbol{t}_{\mathcal{S}})}\big[\log \frac{p_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{x}_{\mathcal{S}}, \boldsymbol{\Omega} \mid \boldsymbol{t}_{\mathcal{S}})}{q_{\boldsymbol{\phi},\boldsymbol{\eta}}(\boldsymbol{\Omega} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})}\big] && (3)\\
&= \mathbb{E}_{q_{\boldsymbol{\phi},\boldsymbol{\theta}}(\boldsymbol{\Omega}|\boldsymbol{x}_{\mathcal{T}},\boldsymbol{t}_{\mathcal{S}})}\Big[\sum_{m=1}^{T} \log p_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{x}_m \mid \boldsymbol{\Omega}, t_m) + \log \frac{\prod_{m=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{\lambda}_m \mid t_m) \prod_{d=1}^{D} p_{\boldsymbol{\eta}}(\boldsymbol{z}_{1:T,d}^{\text{view}} \mid \boldsymbol{\lambda})}{q_{\boldsymbol{\phi}}(\boldsymbol{\lambda} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}}) \prod_{d=1}^{D} q_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{Q},d}^{\text{view}} \mid \boldsymbol{z}_{\mathcal{T},d}^{\text{view}}, \boldsymbol{\lambda}) q_{\boldsymbol{\phi}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})}\\
&\qquad \log \frac{\prod_{k=1}^{K} p(\nu_k) p(z_k^{\text{pres}} \mid \nu_k) p(\boldsymbol{z}_k^{\text{obj}}) p(\boldsymbol{z}^{\text{bck}})}{\prod_{k=1}^{K} q_{\boldsymbol{\phi}}(\nu_k \mid \boldsymbol{x}_{\mathcal{T}}) q_{\boldsymbol{\phi}}(z_k^{\text{pres}} \mid \boldsymbol{x}_{\mathcal{T}}) q_{\boldsymbol{\phi}}(\boldsymbol{z}_k^{\text{obj}} \mid \boldsymbol{x}_{\mathcal{T}}) q_{\boldsymbol{\phi}}(\boldsymbol{z}^{\text{bck}} \mid \boldsymbol{x}_{\mathcal{T}})}\Big] && (4)\\
&= \underbrace{\sum_{m\in\mathcal{T}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{\mathcal{T}}^{\text{view}}|\boldsymbol{x}_{\mathcal{T}},\boldsymbol{t}_{\mathcal{S}}) q_{\boldsymbol{\phi}}(\boldsymbol{\Omega}^{\setminus\text{view}}|\boldsymbol{x}_{\mathcal{T}})}\big[\log p_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{x}_m \mid \boldsymbol{\Omega}_{\mathcal{T}}, t_m)\big]}_{\text{observation reconstruction loss}} && (5)\\
&\quad + \underbrace{\sum_{m\in\mathcal{Q}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\lambda}_{\mathcal{T}}|\boldsymbol{x}_{\mathcal{T}},\boldsymbol{t}_{\mathcal{T}}) q_{\boldsymbol{\phi}}(\boldsymbol{\lambda}_{\mathcal{Q}}|\boldsymbol{\lambda}_{\mathcal{T}},\boldsymbol{t}_{\mathcal{Q}}) q_{\boldsymbol{\phi}}(\boldsymbol{z}_{\mathcal{T}}^{\text{view}}|\boldsymbol{x}_{\mathcal{T}},\boldsymbol{t}_{\mathcal{S}}) q_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{Q}}^{\text{view}}|\boldsymbol{z}_{\mathcal{T}}^{\text{view}},\boldsymbol{\lambda}_{\mathcal{S}}) q_{\boldsymbol{\phi}}(\boldsymbol{\Omega}^{\setminus\text{view}}|\boldsymbol{x}_{\mathcal{T}})}\big[\log p_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{x}_m \mid \boldsymbol{\Omega}, t_m)\big]}_{\text{prediction reconstruction loss}} && (6)\\
&\quad - D_{KL}\big(q_{\boldsymbol{\phi}}(\boldsymbol{\lambda}_{\mathcal{S}} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}}) \| p_{\boldsymbol{\theta}}(\boldsymbol{\lambda}_{\mathcal{S}} \mid \boldsymbol{t}_{\mathcal{S}})\big) && (7)\\
&\quad - \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\lambda}_{\mathcal{S}}|\boldsymbol{x}_{\mathcal{T}},\boldsymbol{t}_{\mathcal{S}})}\big[\sum_{d=1}^{D} D_{KL}\big(q_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{Q},d}^{\text{view}} \mid \boldsymbol{z}_{\mathcal{T},d}^{\text{view}}, \boldsymbol{\lambda}_{\mathcal{S}}) q_{\boldsymbol{\phi}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}}) \| p_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{S},d}^{\text{view}} \mid \boldsymbol{\lambda}_{\mathcal{S}})\big)\big] && (8)\\
&\quad - \sum_{k=1}^{K} D_{KL}(q_{\boldsymbol{\phi}}(\nu_k \mid \boldsymbol{x}_{\mathcal{T}}) \| p(\nu_k)) - \sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\phi}}(\nu_k|\boldsymbol{x}_{\mathcal{T}})}\big[D_{KL}\big(q_{\boldsymbol{\phi}}(z_k^{\text{pres}} \mid \boldsymbol{x}_{\mathcal{T}}) \| p(z_k^{\text{pres}} \mid \nu_k)\big)\big] && (9)\\
&\quad - \sum_{k=1}^{K} D_{KL}\big(q_{\boldsymbol{\phi}}(\boldsymbol{z}_k^{\text{obj}} \mid \boldsymbol{x}_{\mathcal{T}}) \| p(\boldsymbol{z}_k^{\text{obj}})\big) - D_{KL}\big(q_{\boldsymbol{\phi}}(\boldsymbol{z}^{\text{bck}} \mid \boldsymbol{x}_{\mathcal{T}}) \| p(\boldsymbol{z}^{\text{bck}})\big) && (10)
\end{aligned}
$$

Here $\boldsymbol{\Omega}$ is the simplification of all latent variables, i.e., $\boldsymbol{\Omega} = \big\{\boldsymbol{z}^{\text{bck}}, \boldsymbol{z}^{\text{obj}}, \boldsymbol{z}^{\text{pres}}, \boldsymbol{\nu}, \boldsymbol{\lambda}_{\mathcal{S}}, \boldsymbol{z}_{\mathcal{S}}^{\text{view}}\big\}$. Let $\boldsymbol{\Omega}_{\mathcal{T}} =$

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

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

$\{\boldsymbol{z}^{\text{bck}}, \boldsymbol{z}^{\text{obj}}, \boldsymbol{z}^{\text{pres}}, \boldsymbol{\nu}, \boldsymbol{\lambda}_{\mathcal{T}}, \boldsymbol{z}_{\mathcal{Q}}^{\text{view}}\}, \boldsymbol{\Omega}_{\mathcal{T}} = \{\boldsymbol{z}^{\text{bck}}, \boldsymbol{z}^{\text{obj}}, \boldsymbol{z}^{\text{pres}}, \boldsymbol{\nu}, \boldsymbol{\lambda}_{\mathcal{T}}, \boldsymbol{z}_{\mathcal{Q}}^{\text{view}}\}$. Now the observation reconstruction loss and prediction reconstruction loss can be respectively expressed as:

$$\mathcal{L}_{\mathcal{T}} = \sum_{m \in \mathcal{T}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\Omega}_{\mathcal{T}} | \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{T}})} \big[\log p_{\boldsymbol{\theta}, \boldsymbol{\eta}}(\boldsymbol{x}_m \mid \boldsymbol{\Omega}_{\mathcal{T}}, t_m) \big] \tag{11}$$

$$\mathcal{L}_{\mathcal{Q}} = \sum_{m \in \mathcal{Q}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\Omega}_{\mathcal{T}} | \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{T}}) q_{\boldsymbol{\phi}, \boldsymbol{\eta}}(\boldsymbol{\Omega}_{\mathcal{Q}} | \boldsymbol{\Omega}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{Q}})} \big[\log p_{\boldsymbol{\theta}, \boldsymbol{\eta}}(\boldsymbol{x}_m \mid \boldsymbol{\Omega}_{\mathcal{Q}}, \boldsymbol{t}_m) \big] \tag{12}$$

The loss for each item is calculated as:

$$\log p_{\boldsymbol{\theta}, \boldsymbol{\eta}}(\boldsymbol{x}_m \mid \boldsymbol{\Omega}, t_m) = \frac{1}{2\sigma_x^2} \sum_{n=1}^{N} \|\boldsymbol{x}_{m,n} - \sum_{k=0}^{K} \pi_{m,k,n} \boldsymbol{a}_{m,k,n}\|_2^2 + \frac{NC}{2} \log 2\pi\sigma_x^2 \tag{13}$$

$$D_{KL}\big(q_\phi(\boldsymbol{\lambda}_{t,d} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{T}}) \| p_\theta(\boldsymbol{\lambda}_{t,d} \mid \boldsymbol{t}_{\mathcal{T}})\big) = \frac{\|\boldsymbol{\mu}_{t,d}(\boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{T}}) - \boldsymbol{\mu}_{t,d}(\boldsymbol{t}_{\mathcal{T}})\|_2^2}{\sigma_w^2} \tag{14}$$

$$D_{KL}(q_\phi(\nu_k \mid \boldsymbol{x}_{\mathcal{T}}) \| p(\nu_k)) = \log \frac{\Gamma(\tau_{k,1} + \tau_{k,2})}{\Gamma(\tau_{k,1}) \Gamma(\tau_{k,2})} - \log \frac{\alpha}{K} \tag{15}$$

$$+ \left(\tau_{k,1} - \frac{\alpha}{K}\right) \psi(\tau_{k,1}) + (\tau_{k,2} - 1) \psi(\tau_{k,2}) \tag{16}$$

$$- \left(\tau_{k,1} + \tau_{k,2} - \frac{\alpha}{K} - 1\right) \psi(\tau_{k,1} + \tau_{k,2}) \tag{17}$$

$$\mathbb{E}_{q_\phi(\boldsymbol{\nu}_k | \boldsymbol{x}_{\mathcal{T}})} \big[D_{KL}\big(q_\phi(z_k^{\text{pres}} \mid \boldsymbol{x}_{\mathcal{T}}) \| p(z_k^{\text{pres}} \mid \boldsymbol{\nu}_k)\big) \big] = \psi(\tau_{k,1} + \tau_{k,2}) + \kappa_k (\log(\kappa_k) - \psi(\tau_{k,1})) \tag{18}$$

$$+ (1 - \kappa_k)(\log(1 - \kappa_k) - \psi(\tau_{k,2})) \tag{19}$$

$$D_{KL}\big(q_\phi(\boldsymbol{\lambda}_{t,d} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{Q}}) \| p_\theta(\boldsymbol{\lambda}_{t,d} \mid \boldsymbol{t}_{\mathcal{Q}})\big) = \frac{\|\boldsymbol{\mu}_{t,d}(\boldsymbol{x}_{\mathcal{Q}}, \boldsymbol{t}_{\mathcal{Q}}) - \boldsymbol{\mu}_{t,d}\|_2^2}{\sigma_w^2} \tag{20}$$

$$D_{KL}\big(q_\phi(\boldsymbol{z}_{\text{bck}} \mid \boldsymbol{x}_{\mathcal{T}}) \| p(\boldsymbol{z}^{\text{bck}})\big) = \mu^{\text{bck}^2} + \sigma^{\text{bck}^2} - \log \sigma^{\text{bck}^2} - 1 \tag{21}$$

$$D_{KL}\big(q_\phi(\boldsymbol{z}_k^{\text{obj}} \mid \boldsymbol{x}_{\mathcal{T}}) \| p(\boldsymbol{z}_k^{\text{obj}})\big) = \sum_i \left(\mu_{k,i}^{\text{obj}^2} + \sigma_{k,i}^{\text{obj}^2} - \log \sigma_{k,i}^{\text{obj}^2} - 1\right) \tag{22}$$

1.2 KL DIVERGENCE OF VIEWPOINT LATENT VARIABLES

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\lambda} | \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})} \big[\sum_{d=1}^{D} D_{KL}\big(q_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{Q},d}^{\text{view}} \mid \boldsymbol{z}_{\mathcal{T},d}^{\text{view}}, \boldsymbol{\lambda}) q_{\boldsymbol{\phi}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}}) \| p_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{S},d}^{\text{view}} \mid \boldsymbol{\lambda})\big) \big] \tag{23}$$

$$= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\lambda} | \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})} \big[\sum_{d=1}^{D} \mathbb{E}_{q_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{Q},d}^{\text{view}} | \boldsymbol{z}_{\mathcal{T},d}^{\text{view}}, \boldsymbol{\lambda}) q_{\boldsymbol{\phi}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} | \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})} \log \frac{q_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{Q},d}^{\text{view}} \mid \boldsymbol{z}_{\mathcal{T},d}^{\text{view}}, \boldsymbol{\lambda}) q_{\boldsymbol{\phi}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})}{p_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{S},d}^{\text{view}} \mid \boldsymbol{\lambda})} \big] \tag{24}$$

$$= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\lambda} | \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})} \big[\sum_{d=1}^{D} \mathbb{E}_{q_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{Q},d}^{\text{view}} | \boldsymbol{z}_{\mathcal{T},d}^{\text{view}}, \boldsymbol{\lambda}) q_{\boldsymbol{\phi}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} | \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})} \log \frac{q_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{Q},d}^{\text{view}} \mid \boldsymbol{z}_{\mathcal{T},d}^{\text{view}}, \boldsymbol{\lambda}) q_{\boldsymbol{\phi}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})}{p_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{Q},d}^{\text{view}} \mid \boldsymbol{z}_{\mathcal{T},d}^{\text{view}}, \boldsymbol{\lambda}) p_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} \mid \boldsymbol{\lambda})} \big] \tag{25}$$

$$\text{// given } \boldsymbol{\lambda} \sim q_\phi(\boldsymbol{\lambda} \mid \boldsymbol{x}_{\mathcal{T}}), \boldsymbol{z}_{\mathcal{T},d}^{\text{view}} \sim q_\phi(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}}), \text{ then } q_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{Q},d}^{\text{view}} \mid \boldsymbol{z}_{\mathcal{T},d}^{\text{view}}, \boldsymbol{\lambda}) = p_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{Q},d}^{\text{view}} \mid \boldsymbol{z}_{\mathcal{T},d}^{\text{view}}, \boldsymbol{\lambda}) \tag{26}$$

$$= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\lambda} | \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})} \big[\sum_{d=1}^{D} \mathbb{E}_{q_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{Q},d}^{\text{view}} | \boldsymbol{z}_{\mathcal{T},d}^{\text{view}}, \boldsymbol{\lambda}) q_{\boldsymbol{\phi}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} | \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})} \log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})}{p_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} \mid \boldsymbol{\lambda})} \big] \tag{27}$$

$$= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\lambda} | \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})} \big[\sum_{d=1}^{D} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} | \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})} \log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})}{p_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} \mid \boldsymbol{\lambda})} \big] \tag{28}$$

$$= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\lambda} | \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})} \Big[\sum_{d=1}^{D} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} | \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}})} \big[\log q_{\boldsymbol{\phi}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{S}}) - \log p_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} \mid \boldsymbol{\lambda}) \big] \Big] \tag{29}$$

where $p_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}} \mid \boldsymbol{\lambda}) = \frac{1}{(2\pi)^{\frac{|\mathcal{T}|}{2}}} \frac{1}{|\boldsymbol{C}_{\boldsymbol{\eta}}(\mathcal{T})|^{\frac{1}{2}}} \exp\big\{ -\frac{1}{2} (\boldsymbol{z}_{\mathcal{T},d}^{\text{view}})^\top \boldsymbol{C}_{\boldsymbol{\eta}}^{-1}(\mathcal{T}) \boldsymbol{z}_{\mathcal{T},d}^{\text{view}} \big\}$, see , in Eq 38.

2 DETAILS OF GENERATION

Suppose each *static* visual scene is composed of T consecutive viewpoints and each instance (video) is independently and identically distributed. Hereby, for simple description, the following will take a single video instance as an example. Let N and C respectively denote the number of pixels and channels in each frame $\boldsymbol{x}_t$ $(1 \leq t \leq T)$ of the video, and K denotes the maximum number of objects that appear in the visual scene. Each pixel $\boldsymbol{x}_{t,n}$ of each frame is a weighted summation of $K+1$ components at that pixel, with K describing the objects and one describing the background. $K+1$ layers of the image correspond to the $K+1$ components composed of N pixels. In the compositional modeling, layers of each frame $\boldsymbol{x}_t$ consist of pixel-wise weights $\boldsymbol{\pi}_t \in \mathbb{R}^{(K+1)\times N}$ and the expected pixel-wise RGB value $\boldsymbol{a}_t \in \mathbb{R}^{(K+1)\times N \times C}$. Both are generated by some representations (including latent variables, deterministic values, neural networks, etc.). We will express them below.

View-independent representations. we define a set of object-centric latent variables of $K+1$ entities from T viewpoints that describes the 3D visual scene, including $\boldsymbol{z}^{\text{obj}}, \boldsymbol{z}^{\text{bck}}, \boldsymbol{z}^{\text{pres}}, \boldsymbol{\nu}$.

- $\boldsymbol{z}^{\text{obj}} = \boldsymbol{z}^{\text{obj}}_{1:K}$ describes the 3D view-independent representations of objects. View-independent means the physical attributes of objects (such as shape, appearance, etc.) keep constant under different multiple viewpoints. $\boldsymbol{z}^{\text{obj}}_k$ $(1 \leq k \leq K)$ is independently and identically distributed.
- $\boldsymbol{z}^{\text{bck}}$ denotes the latent representation of the background appearance. We need not represent the shape of the background because the corresponding complete shape is 1.
- $\boldsymbol{z}^{\text{pres}} = \boldsymbol{z}^{\text{pres}}_{1:K}, \boldsymbol{\nu} = \boldsymbol{\nu}_{1:K}$ denote the latent variables that indicate the presence of objects. The advantage of using the latents is that the uncertain number of objects in different visual scenes can be added up. $z^{\text{pres}}_k(1 \leq k \leq K)$ denotes whether object k appears in a visual scene, following a Bernoulli distribution. The parameter of the distribution is controlled by latent variable ν_k that follows the conjugate prior, i.e. $z^{\text{pres}}_k \sim \text{Bernoulli}(\nu_k), \nu_k \sim \text{Beta}(\alpha/K, 1)$, where α is the hyperparameter, K denotes the object numbers.

View-dependent representations. Different from previous works [Li et al., 2020, Chen et al., 2021], we learn view representations through finding the relationship between frames, rather than directly leveraging viewpoint labels. Meanwhile, the view correlation based on temporal modeling can motivate the model to predict the novel scenes unseen given any time. The related view-dependent representations include $\boldsymbol{\lambda}, \boldsymbol{z}^{\text{view}}$.

- $\boldsymbol{\lambda} \in \mathbb{R}^{T\times D\times D_\lambda}$ represents the spatial latent variable that reflects the position characteristics of the camera under different frames, where D is the dimension of the view latent (i.e. $\boldsymbol{z}^{\text{view}}$) corresponding to the frame. D_λ is the dimension of the spatial representation that influences the meanings of each dimension in the view latent. $\boldsymbol{\lambda}_t$ potentially affects the change of viewpoints ((e.g. the distance, height, rotation of the camera) at time t. $\boldsymbol{\lambda}_{t,d}$ $(1 \leq t \leq T, 1 \leq d \leq D)$ is distributed in a linear subspace.
- $\boldsymbol{z}^{\text{view}} = \boldsymbol{z}^{\text{view}}_{1:T} \in \mathbb{R}^{T\times D}$ denotes the view latent variables. Videos perform in a way that the closer the distance of two frames, the smaller the difference between the corresponding viewpoint information, and the bigger on the contrary. To build the correlation, we define $\boldsymbol{z}^{\text{view}}$ as a Gaussian process (GP) prior parameterized by the spatial latent variable $\boldsymbol{\lambda}$.

Additional Notations. In addition to the latent variables defined above, we also need some non-latent notations to generate T frames, including $\boldsymbol{s}^{\text{shp}}$, $\boldsymbol{o}$, $\boldsymbol{\pi}$, $\boldsymbol{a}$.

- $\boldsymbol{s}^{\text{shp}} \in [0,1]^{T\times K\times N}$ describes the complete shape of different objects at different time t. $\boldsymbol{s}^{\text{shp}}_{t,\cdot,k}$ ($\cdot$ represents all indexes are selected) represents the complete shape of the kth object in the 2D image corresponding to the tth frame. The range of [0, 1] guarantees the subsequent rationality of processing the occlusion. Since the complete shape of the background is a constant of 1, values of $\boldsymbol{s}^{\text{shp}}_{t,\cdot,k}$ can be computed by the neural network g_{shp} with $\boldsymbol{z}^{\text{obj}}_k$ and $\boldsymbol{z}^{\text{view}}_t$ as inputs followed by a sigmoid activation and $\boldsymbol{z}^{\text{bck}}$ need not participate in the computation.
- $\boldsymbol{o} \in \mathbb{R}^{T\times K}$ describes the occlusion order of different objects in the different frame. $o_{t,k}$ denotes the order of the kth object under the projected 2D image at the tth frame. $\boldsymbol{o}$ is obtained by the neural network g_{ord} with $\boldsymbol{z}^{\text{obj}}_k$ and $\boldsymbol{z}^{\text{view}}_t$ as inputs since the occlusion order of the same object varies at different viewpoints.
- $\boldsymbol{\pi} \in [0,1]^{T\times(K+1)\times N}$ represents the pixel-wise weights of each layer, i.e. geometrically represents the observed shape of each object. The $\boldsymbol{\pi}$ here is different from $\boldsymbol{s}^{\text{shp}}$ in that the shape of an object may be partially observed or completely invisible due to partial or complete occlusion. $K+1$ observed shapes of the nth pixel at frame t satisfy $\sum_{k=0}^{K} \pi_{t,k,n} = 1$ $(1 \leq t \leq T, 1 \leq n \leq N)$.

- $\boldsymbol{a} \in \mathbb{R}^{T \times (K+1) \times N \times C}$ describes the complete appearance of all entities (objects or backgrounds). $\boldsymbol{a}_{t,k,n}$ is numerically equivalent to the expected RGB value of component k at the nth pixel of the tth frame. The background appearance $\boldsymbol{a}_{t,k}$ $(k = 0)$ is achieved by the neural network $g_{\text{apc}}^{\text{bck}}$ with $\boldsymbol{z}_t^{\text{view}}$ and $\boldsymbol{z}^{\text{bck}}$ as inputs, meanwhile the kth object appearance $\boldsymbol{a}_{t,k}$ $(1 \leq k \leq K)$ is achieved by another neural network $g_{\text{apc}}^{\text{obj}}$ with $\boldsymbol{z}_t^{\text{view}}$ and $\boldsymbol{z}_k^{\text{obj}}$ as inputs.

Likelihood Function. After generating the observed shapes $\boldsymbol{\pi}$ and appearance $\boldsymbol{a}$ of each layer, we can use a weighted summation of each layer to reconstruct the image. Its likelihood is expressed as:

$$\log p(\boldsymbol{x}_{1:T} \mid \boldsymbol{\pi}, \boldsymbol{a}) = \sum_{t=1}^{T} \sum_{n=1}^{N} \log \mathcal{N}(\underbrace{\pi_{t,0,n} \cdot \boldsymbol{a}_{t,0,n}}_{\text{Background}} + \sum_{k=1}^{K} \underbrace{\pi_{t,k,n} \cdot \boldsymbol{a}_{t,k,n}}_{\text{Objects}}, \hat{\sigma}_x^2 \boldsymbol{I}) \tag{30}$$

where $\hat{\sigma}_x$ is the hyperparameter. The style of the likelihood function is similar to Slot Attention [Locatello et al., 2020] in order to improve the reconstruction.

3 DETAILS OF INFERENCE

3.1 APPROXIMATION OF PREDICTED SPATIAL LATENT VARIABLES

$\boldsymbol{\lambda}_{t,d} \in \mathbb{R}^{D_\lambda} (1 \leq t \leq |\mathcal{S}|, 1 \leq d \leq D)$ denotes the spatial latent representations corresponding to $\boldsymbol{z}_{t,d}^{\text{view}}$. In the generative process, $\boldsymbol{\lambda}_{t,d}$ is distributed in a linear subspace, i.e.,

$$\boldsymbol{\lambda}_{t,d} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{w}_t, \sigma_w^2 \boldsymbol{I}), \boldsymbol{A} \in \mathbb{R}^{D_\lambda \times |\boldsymbol{w}|} \tag{31}$$

where $\boldsymbol{A}$ and σ_w^2 are the hyperparameters. For the posterior of $\boldsymbol{\lambda}$, we can simply define the distribution on $\boldsymbol{\lambda}_{\mathcal{T}}$ that satisfies the linear distribution:

$$q(\boldsymbol{\lambda}_{t,d} \mid \boldsymbol{x}_t, \boldsymbol{w}_t) \sim \mathcal{N}(\boldsymbol{\mu}_d(\boldsymbol{x}_t, \boldsymbol{w}_t), \sigma_w^2 \boldsymbol{I}), \quad t \in \mathcal{T}, 1 \leq d \leq D \tag{32}$$

$q(\boldsymbol{\lambda}_{t,d} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{w}_{\mathcal{S}})$ for $t \in \mathcal{Q}$ is difficult. We apply the Least Square Error to find the optimal mean curve that satisfies a linear relationship w.r.t. $\boldsymbol{w}_t$:

$$\hat{\boldsymbol{A}}_d^* = \underset{\hat{\boldsymbol{A}}_d}{\arg\min} \left\| \boldsymbol{\Phi}_d - \boldsymbol{W}_{\mathcal{T}} \hat{\boldsymbol{A}}_d^\top \right\|_2^2 \quad , \hat{\boldsymbol{A}}_d \in \mathbb{R}^{D_\lambda \times |\boldsymbol{w}|} \tag{33}$$

where $\boldsymbol{\Phi}_d = \left[\boldsymbol{\mu}_{1,d}, ..., \boldsymbol{\mu}_{|\mathcal{T}|,d}\right]^\top \in \mathbb{R}^{|\mathcal{T}| \times D_\lambda}$, $\boldsymbol{W}_{\mathcal{T}} = \left[\boldsymbol{w}_1, ..., \boldsymbol{w}_{|\mathcal{T}|}\right]^\top \in \mathbb{R}^{|\mathcal{T}| \times |\boldsymbol{w}|}$. $\hat{\boldsymbol{A}}_d^*$ can be analytically solved and the optimal $\hat{\boldsymbol{A}}_d^*$ is described as:

$$\hat{\boldsymbol{A}}_d^* = \boldsymbol{\Phi}_d^\top \boldsymbol{W}_{\mathcal{T}} (\boldsymbol{W}_{\mathcal{T}}^\top \boldsymbol{W}_{\mathcal{T}})^{-1} \tag{34}$$

Then $q(\boldsymbol{\lambda}_{t,d} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{w}_{\mathcal{S}})$ for $t \in \mathcal{Q}$ can be approximated as:

$$q(\boldsymbol{\lambda}_{t,d} \mid \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{w}_{\mathcal{S}}) = \mathcal{N}(\hat{\boldsymbol{A}}_d^* \boldsymbol{w}_t, \sigma_w^2 \boldsymbol{I}) \tag{35}$$

3.2 GAUSSIAN PROCESSES AND INFERENCE OF PREDICTED VIEW LATENT REPRESENTAIONS

if the variable $\boldsymbol{z}_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$ satisfies the Gaussian Processes (GPs):

$$p(\boldsymbol{z}_S \mid \boldsymbol{\lambda}) \sim N\Bigg(\boldsymbol{0}, \begin{bmatrix} \kappa_\eta(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_1) & \cdots & \kappa_\eta(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_{|\mathcal{S}|}) \\ \vdots & \ddots & \vdots \\ \kappa_\eta(\boldsymbol{\lambda}_{|\mathcal{S}|}, \boldsymbol{\lambda}_1) & \cdots & \kappa_\eta(\boldsymbol{\lambda}_{|\mathcal{S}|}, \boldsymbol{\lambda}_{|\mathcal{S}|}) \end{bmatrix}\Bigg) \tag{36}$$

To simplify the analysis, we randomly divides the covariance matrix to the $\boldsymbol{z}_{\mathcal{Q}}$-dependent sub-matrix and $\boldsymbol{z}_{\mathcal{Q}}$-independent sub-matrix (aggregate different subsets together by translation).

$$\begin{bmatrix} \kappa_\eta(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_1) & \cdots & \kappa_\eta(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_{|\mathcal{S}|}) \\ \vdots & \ddots & \vdots \\ \kappa_\eta(\boldsymbol{\lambda}_{|\mathcal{S}|}, \boldsymbol{\lambda}_1) & \cdots & \kappa_\eta(\boldsymbol{\lambda}_{|\mathcal{S}|}, \boldsymbol{\lambda}_{|\mathcal{S}|}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{C}_\eta(\mathcal{T}) & \boldsymbol{R}_\eta(\mathcal{T}, \mathcal{Q}) \\ \boldsymbol{R}_\eta(\mathcal{Q}, \mathcal{T}) & \boldsymbol{C}_\eta(\mathcal{Q}) \end{bmatrix} \tag{37}$$

where

$$\boldsymbol{C}_{\eta}(\mathcal{H}) = \begin{bmatrix} \kappa_{\eta}\left(\boldsymbol{\lambda}_i, \boldsymbol{\lambda}_i\right) & \cdots & \kappa_{\eta}\left(\boldsymbol{\lambda}_i, \boldsymbol{\lambda}_{i+j}\right) \\ \vdots & \ddots & \vdots \\ \kappa_{\eta}\left(\boldsymbol{\lambda}_{i+j}, \boldsymbol{\lambda}_i\right) & \cdots & \kappa_{\eta}\left(\boldsymbol{\lambda}_{i+j}, \boldsymbol{\lambda}_{i+j}\right) \end{bmatrix}, \tag{38}$$

$$\boldsymbol{R}_{\eta}(\mathcal{X}, \mathcal{Y}) = \begin{bmatrix} \kappa_{\eta}\left(\boldsymbol{\lambda}_u, \boldsymbol{\lambda}_v\right) & \cdots & \kappa_{\eta}\left(\boldsymbol{\lambda}_u, \boldsymbol{\lambda}_{v+n}\right) \\ \vdots & \ddots & \vdots \\ \kappa_{\eta}\left(\boldsymbol{\lambda}_{u+m}, \boldsymbol{\lambda}_v\right) & \cdots & \kappa_{\eta}\left(\boldsymbol{\lambda}_{u+m}, \boldsymbol{\lambda}_{v+n}\right) \end{bmatrix} \tag{39}$$

where $i \sim i+j \in \mathcal{H}, \mathcal{H} = \{\mathcal{T}, \mathcal{Q}\}; u \sim u+m \in \mathcal{X}, v \sim v+n \in \mathcal{Y}, \mathcal{X} \neq \mathcal{Y}, \mathcal{X}, \mathcal{Y} \in \{\mathcal{T}, \mathcal{Q}\}$.

given the observation set $\boldsymbol{z}_{\mathcal{T}}$ and $\boldsymbol{\lambda}_{\mathcal{S}}$, $p(\boldsymbol{z}_{\mathcal{Q}} \mid \boldsymbol{z}_{\mathcal{T}}, \boldsymbol{\lambda})$ can be calculated analytically using properties of the multivariate Gaussian distribution [Bishop and Nasrabadi, 2006]:

$$p(\boldsymbol{z}_{\mathcal{Q}} \mid \boldsymbol{z}_{\mathcal{T}}, \boldsymbol{\lambda}) = \mathcal{N}\big(\boldsymbol{\mu}_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{T}}, \boldsymbol{\lambda}), \boldsymbol{\Sigma}_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{T}}, \boldsymbol{\lambda})\big) \tag{40}$$

$$\boldsymbol{\mu}_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{T}}, \boldsymbol{\lambda}) = \boldsymbol{R}_{\boldsymbol{\eta}}(\mathcal{Q}, \mathcal{T}) \boldsymbol{C}_{\boldsymbol{\eta}}^{-1}(\mathcal{T}) \boldsymbol{z}_{\mathcal{T}} \tag{41}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{T}}, \boldsymbol{\lambda}) = \boldsymbol{C}_{\boldsymbol{\eta}}(\mathcal{Q}) - \boldsymbol{R}_{\boldsymbol{\eta}}(\mathcal{Q}, \mathcal{T}) \boldsymbol{C}_{\boldsymbol{\eta}}^{-1}(\mathcal{T}) \boldsymbol{R}_{\boldsymbol{\eta}}(\mathcal{Q}, \mathcal{T})^{\top} \tag{42}$$

where $\boldsymbol{C}_{\boldsymbol{\eta}}(\mathcal{T}) \in \mathbb{R}^{|\mathcal{T}| \times |\mathcal{T}|}, \boldsymbol{R}_{\boldsymbol{\eta}}(\mathcal{Q}, \mathcal{T}) \in \mathbb{R}^{|\mathcal{Q}| \times |\mathcal{T}|}, \boldsymbol{C}_{\boldsymbol{\eta}}(\mathcal{Q}) \in \mathbb{R}^{|\mathcal{Q}| \times |\mathcal{Q}|}$.

According to the derivation above, $q(\boldsymbol{z}_{\mathcal{Q},d}^{\text{view}} \mid \boldsymbol{z}_{\mathcal{T},d}^{\text{view}}, \boldsymbol{\lambda})$ can be analytically rsampled from Eq 40, i.e.

$$\boldsymbol{z}_{\mathcal{Q},d}^{\text{view}} = \boldsymbol{\mu}_{\boldsymbol{\eta}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}}, \boldsymbol{\lambda}) + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{\frac{1}{2}}(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}}, \boldsymbol{\lambda})\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \in \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{43}$$

$$\boldsymbol{z}_{\mathcal{Q}}^{\text{view}} = \text{concatenate}(\boldsymbol{z}_{\mathcal{T},1}^{\text{view}}, \boldsymbol{z}_{\mathcal{T},2}^{\text{view}}, \cdots, \boldsymbol{z}_{\mathcal{T},D}^{\text{view}}, \text{axis} = \text{“view dim”}), \quad \boldsymbol{z}_{\mathcal{Q}}^{\text{view}} \in \mathbb{R}^{|\mathcal{Q}| \times D} \tag{44}$$

3.3 MATHEMATICAL FORM OF INFERENCE

We detail the algorithm of inference in this section. algorithm 1 describes the whole mathematical form. It‘s worth mentioning that the view-independent latent variables and view-independent latent variables are inferred based on different nerual networks. More specifically speaking, the feature for view-dependent latent variables is extracted by the neural netowork f_{view} and then the feature will enter into the Transformer to obtain the $\boldsymbol{\lambda}_{\mathcal{T}}$ and $\boldsymbol{z}_{\mathcal{T}}^{\text{view}}$. The feature for view-independent latent variables is extracted by the neural network f_{sa} and enters into the sequential extension of Slot Attention [Locatello et al., 2020]. During the interation in Slot Attention, the view feature $\boldsymbol{y}_{\mathcal{T}}^{\text{view}}$ from the Transoformer will enter into the Slot Attention module, and then concatenate with $\boldsymbol{y}^{\text{attr}}$ initialized with the Gaussian distribution one by one. Note that $\boldsymbol{y}_{\mathcal{T}}^{\text{view}}$ will not be updated during the iteration. different from $\boldsymbol{y}_{\mathcal{T}}^{\text{view}}$, the module will execute the temporal mean of $\boldsymbol{y}^{\text{attr}}$ at each iteration after the cross-attention.

4 DATASETS

The datasets (CLEVR-SIMPLE, CLEVR-COMPLEX, SHOP-SIMPLE, SHOP-COMPLEX) used in this paper are modified based on the official code of CLEVR [Johnson et al., 2017] and SHOP [Nazarczuk and Mikolajczyk, 2020]. More specifically speaking, we have made some improvements to the official code of CLEVR dataset and SHOP dataset, that is, polar coordinates are used to assign a shot position (x_t, y_t, z_t) to each frame of the video. In the polar coordinates, ρ $(\rho > 0)$ represents the radius of the object in a 3D sphere, ϕ $(0 \leq \phi \leq \frac{\pi}{2})$ describes the angle between the object and the z positve half axis, and θ $(0 \leq \theta \leq 2\pi)$ describes the angle between the object and the xy axis. The function of camera coordinates (x_t, y_t, z_t) with respect to time t can be described as:

$$\begin{aligned} x &= \rho \sin\phi \cos\theta \\ y &= \rho \sin\phi \sin\theta \\ z &= \rho \cos\phi \end{aligned}$$

Algorithm 1 Inference of Latent Variables

```
Requires: observed images $\boldsymbol{x}_{\mathcal{T}}$, timesteps $\boldsymbol{t}_{\mathcal{S}} = (\boldsymbol{t}_{\mathcal{T}}, \boldsymbol{t}_{\mathcal{Q}})$, maximum iterations $M_s$.
// extract the feature $\boldsymbol{y}_{\mathcal{T}}^{\text{feat}} \in \mathbb{R}^{|O| \times L \times C}$,$L$ is the product of the height and width corresponding to the feature map
// $\boldsymbol{y}_{\mathcal{T}}^{\text{sa}} \in \mathbb{R}^{|\mathcal{T}| \times N \times D'}$ is another feature map with the neural network $f_{\text{sa}}$
$\boldsymbol{y}_{\mathcal{T}}^{\text{feat}} = f_{\text{feat}}(\boldsymbol{x}_{\mathcal{T}}), \boldsymbol{y}_{\mathcal{T}}^{\text{sa}} = f_{\text{sa}}(\boldsymbol{x}_{\mathcal{T}}), [\boldsymbol{y}_1^{\text{sa}}, \boldsymbol{y}_2^{\text{sa}}, ..., \boldsymbol{y}_{|\mathcal{T}|}^{\text{sa}}] = \text{split}(\boldsymbol{y}_{\mathcal{T}}^{\text{sa}}, \text{axis} = 0)$
$\boldsymbol{y}_{O}^{\text{feat}} = \text{MultiHeadSelfAttention}\big(\text{3DPositionEmbedding}(\boldsymbol{y}_{O}^{\text{feat}})\big)$
$\boldsymbol{y}_{\mathcal{T}}^{\text{view}} = \text{mean}(\boldsymbol{y}_{\mathcal{T}}^{\text{feat}}, \text{axis} = 1); \quad [\boldsymbol{y}_1^{\text{view}}, \boldsymbol{y}_2^{\text{view}}, ..., \boldsymbol{y}_{|\mathcal{T}|}^{\text{view}}] = \text{split}(\boldsymbol{y}_{\mathcal{T}}^{\text{view}}, \text{axis} = 1)$
// do the spatial mean on $\boldsymbol{y}_{\mathcal{T}}^{\text{feat}}$, and then encode to the posterior of $\boldsymbol{\lambda}_{\mathcal{S}} = (\boldsymbol{\lambda}_{\mathcal{T}}, \boldsymbol{\lambda}_{\mathcal{Q}})$, where $\boldsymbol{\mu}_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}| \times D \times D_{\lambda}}$
$\boldsymbol{y}_{\mathcal{T}}^{\text{feat}} = \text{Downsample}(\text{MultiHeadSelfAttention}(\boldsymbol{y}_{\mathcal{T}}^{\text{feat}}))$
$\boldsymbol{y}_{\mathcal{T}}^{\lambda} = \text{mean}(\boldsymbol{y}_{\mathcal{T}}^{\text{feat}}, \text{axis} = 1)$
$\boldsymbol{w}_{\mathcal{S}} = \text{TimestepEncoding}(\boldsymbol{t}_{\mathcal{S}}), \boldsymbol{w}_{\mathcal{S}} = (\boldsymbol{w}_{\mathcal{T}}, \boldsymbol{w}_{\mathcal{Q}})$
// $\hat{\boldsymbol{A}}_d^*(1 \leq d \leq D)$ can be obtained by eq 34
$\boldsymbol{\mu}_{\mathcal{T}} = f_{\phi}^{\lambda}(\boldsymbol{y}_{\mathcal{T}}^{\lambda}, \boldsymbol{w}_{\mathcal{T}}), \boldsymbol{\mu}_{\mathcal{Q}} = \text{concatenate}(\big[\hat{\boldsymbol{A}}_1^* \boldsymbol{w}_{\mathcal{Q}}, ..., \hat{\boldsymbol{A}}_D^* \boldsymbol{w}_{\mathcal{Q}}\big], \text{axis} = \text{“view dim”}), \boldsymbol{\mu}_{\mathcal{S}} = [\boldsymbol{\mu}_{\mathcal{T}}, \boldsymbol{\mu}_{\mathcal{Q}}]$
$\boldsymbol{\lambda}_S \sim N(\boldsymbol{\mu}_S, \sigma_{\boldsymbol{w}}^2 \boldsymbol{I})$
$\boldsymbol{y}_k^{\text{obj}} \sim N(\hat{\boldsymbol{\mu}}^{\text{obj}}, \hat{\sigma}^{\text{obj}} \boldsymbol{I}), \quad \forall 1 \leq k \leq K$
$\boldsymbol{y}^{\text{bck}} \sim N(\hat{\boldsymbol{\mu}}^{\text{bck}}, \hat{\sigma}^{\text{bck}} \boldsymbol{I})$
$\boldsymbol{y}_{1:K+1}^{\text{attr}} = [\boldsymbol{y}_{1:K}^{\text{obj}}, \boldsymbol{y}^{\text{bck}}]$,
// do the iteration of sequential Slot Attention
for $s = 1$ to $M$ do $\big\{\forall t \in \mathcal{T}, \forall 1 \leq k \leq K + 1\big\}$
    $\boldsymbol{y}_{t,k}^{\text{full}} = [\boldsymbol{y}_t^{\text{view}}, \boldsymbol{y}_k^{\text{attr}}]$
    $\boldsymbol{a}_t = \underset{K+1}{\text{softmax}}\Big(\frac{k(\boldsymbol{y}_t^{\text{sa}}) \cdot q(\boldsymbol{y}_{t,1:K+1}^{\text{full}})^{\top}}{\sqrt{D_f}}\Big) \in \mathbb{R}^{N \times K}$
    $\boldsymbol{u}_t = \sum_N \underset{N}{\text{softmax}}\Big(\log \boldsymbol{a}_{t,n}\Big) \cdot v(\boldsymbol{y}_{t,n}^{\text{sa}}) \in \mathbb{R}^{K \times D_f}$
    $\hat{\boldsymbol{y}}_{t,k}^{\text{full}} = \text{GRU}(\boldsymbol{y}_{t,k}^{\text{full}}, \boldsymbol{u}_{t,k}), \quad \big[\hat{\boldsymbol{y}}_{t,k}^{\text{view}}, \hat{\boldsymbol{y}}_{t,k}^{\text{attr}}\big] \overset{\text{split}}{\leftarrow} \hat{\boldsymbol{y}}_{t,k}^{\text{full}}$
    $\boldsymbol{y}_k^{\text{attr}} = \text{mean}_{|\mathcal{T}|}\Big(\hat{\boldsymbol{y}}_{1:|\mathcal{T}|,k}^{\text{attr}}\Big)$
end for
$[\boldsymbol{y}_1^{\text{obj}}, ..., \boldsymbol{y}_K^{\text{obj}}, \boldsymbol{y}^{\text{bck}}] = \text{split}(\boldsymbol{y}^{\text{attr}}, \text{axis} = 0)$
// independently and identically sample $\boldsymbol{z}_t^{\text{view}}$ for $1 \leq t \leq |\mathcal{T}|$, then infer $\boldsymbol{z}_{\mathcal{Q}}^{\text{view}}$ for $1 \leq d \leq D$
$\boldsymbol{\mu}_t^{\text{view}}, \boldsymbol{\sigma}_t^{\text{view}} = f_{\phi}^{\text{view}}(\boldsymbol{y}_t^{\text{view}})$
$\boldsymbol{z}_{\mathcal{T}}^{\text{view}} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathcal{T}}^{\text{view}}, \text{diag}(\boldsymbol{\sigma}_{\mathcal{T}}^{\text{view}})^2)$
// $\boldsymbol{\mu}_{\eta}$ and $\boldsymbol{\Sigma}_{\eta}$ can be analytically computed, see eq 40
$\boldsymbol{z}_{\mathcal{Q},d}^{\text{view}} = \boldsymbol{\mu}_{\eta}\left(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}}, \boldsymbol{\lambda}_{\mathcal{S}}\right) + \boldsymbol{\Sigma}_{\eta}^{\frac{1}{2}}\left(\boldsymbol{z}_{\mathcal{T},d}^{\text{view}}, \boldsymbol{\lambda}_{\mathcal{S}}\right)\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
$\boldsymbol{z}_{\mathcal{Q}}^{\text{view}} = \text{concatenate}(\boldsymbol{z}_{\mathcal{Q},1}^{\text{view}}, \boldsymbol{z}_{\mathcal{Q},2}^{\text{view}}, ..., \boldsymbol{z}_{\mathcal{Q},D}^{\text{view}}, \text{axis} = \text{“view dim”})$
$\boldsymbol{z}_S^{\text{view}} = [\boldsymbol{z}_{\mathcal{T}}^{\text{view}}, \boldsymbol{z}_{\mathcal{Q}}^{\text{view}}]$
// infer the view-independent latent variables for $1 \leq k \leq K$
$\boldsymbol{\mu}^{\text{bck}}, \boldsymbol{\sigma}^{\text{bck}} = f_{\phi}^{\text{bck}}(\boldsymbol{y}^{\text{bck}})$
$\boldsymbol{\mu}_k^{\text{obj}}, \boldsymbol{\sigma}_k^{\text{obj}}, \boldsymbol{\tau}_k, \kappa_k = f_{\phi}^{\text{obj}}(\boldsymbol{y}_k^{\text{obj}})$
return $\boldsymbol{\lambda}_{\mathcal{S}}, \boldsymbol{z}_{\mathcal{S}}^{\text{view}}, \boldsymbol{\mu}^{\text{bck}}, \boldsymbol{\sigma}^{\text{bck}}, \boldsymbol{\mu}_{1:K}^{\text{obj}}, \boldsymbol{\sigma}_{1:K}^{\text{obj}}, \boldsymbol{\tau}_{1;K}, \boldsymbol{\kappa}_{1:K}$
```

When constructing the dataset, the polar coordinate configuration corresponding to the camera position of each scene (10 frames) is $\rho \sim U(\rho_{\text{min}}, \rho_{\text{max}}), \phi \sim U(\phi_{\text{min}}, \phi_{\text{max}}), \theta = \frac{2\pi}{10}t(0 \leq t \leq 9)$. In the image rendering process, we remove the code to check whether an object is visible (that is, whether the number of observation pixels of an object reaches the maximum threshold), so that we hope the model can retrieve the occluded or completely occluded objects from the frame relationship. The size of the generated image of CLEVR and SHOP is 108×64. We crop the image to 64×64 (the upper boundary is 10, the lower boundary is 74, the left boundary is 22, and the right boundary is 86).

The CLEVR is further divided into two categories: CLEVR-SIPLE and CLEVR-COMPLEX. CLEVR-SIMPLE includes 3 object categories (with intra class differences), while CLEVR-COMPLEX includes 10 object categories (with intra class differences). Compared with CLEVR-SIMPLE, CLEVR-COMPLEX has greater challenges. SHOP is further divided into two types: SHOP-SIMPlE and SHOP-COMPlEX. SHOP-SIMPlE includes 10 object categories, and the background is

Datasets	CLEVR-SIMPLE				CLEVR-COMPLEX			
Split	Train	Valid	Test	General	Train	Vaid	Test	General
# of Images	5000	100	100	100	5000	100	100	100
# of Objects	3∼6	3∼6	3∼6	7∼10	3∼6	3∼6	3∼6	7∼ 10
# of Views	10				10			
# of Categories	3				10			
# of Backgrounds	1				1			
Image Size	64×64				64×64			
Azimuth θ Elevation ρ Distance ϕ	[0,2π] [10.5,12] [0.15π,0.3π]				[0,2π] [10.5,12] [0.15π,0.3π]			

Table 1: configuration of CLEVR

Datasets	SHOP-SIMPLE				SHOP-COMPLEX			
Split	Train	Valid	Test	General	Train	Vaid	Test	General
# of Images	5000	100	100	100	5000	100	100	100
# of Objects	3∼6	3∼6	3∼6	7∼10	3∼6	3∼6	3∼6	7∼ 10
# of Views	10				10			
# of Categories	multicolumn4cl10	10						
# of Backgrounds	1				2			
Image Size	64×64				64×64			
Azimuth θ Elevation ρ Distance ϕ	[0,2π] [10.5,12] [0.15π,0.3π]				[0,2π] [10.5,12] [0.15π,0.3π]			

Table 2: configuration of SHOP

selected as marble background. Compared with CLEVR, its objects have greater challenges in texture and material. At the same time, the color of some objects is highly similar to the background, which makes it more difficult to identify. SHOP-COMPLEX has two background options. The second one is a brown background, whose color is highly similar to the object color, further improving the recognition difficulty. The detailed configuration can be found in Table 1 and 2. Figure 1 and 2 demonstrate the samples in test sets and general sets of four datasets. It can be seen that the number of objects in the general set is larger than the test sets, and correspondingly the occlusion rate is higher, leading to a more difficult inference.

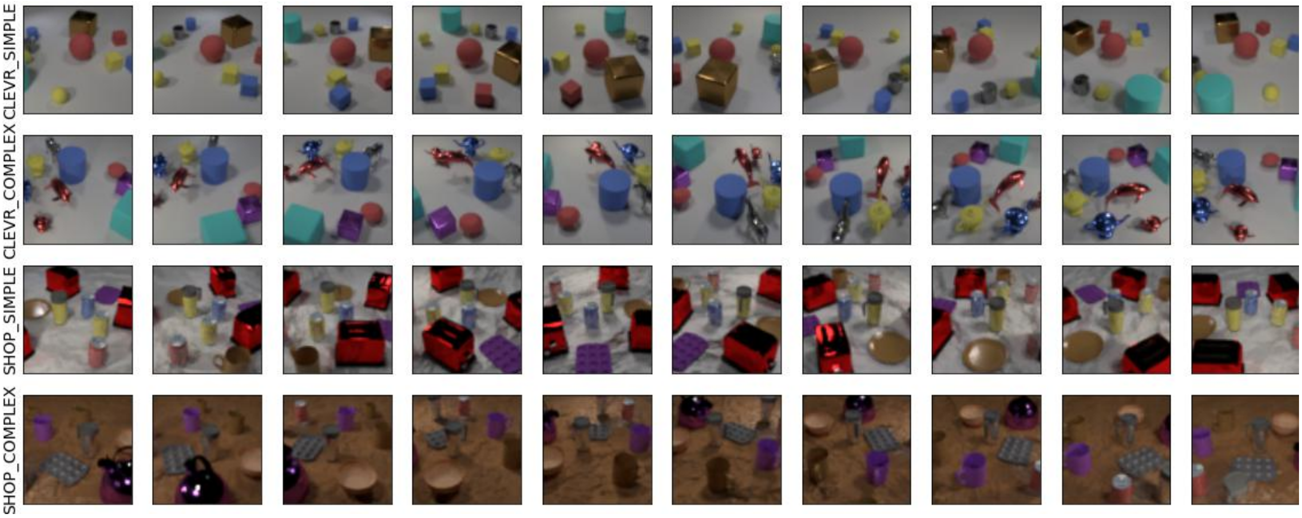


Figure 2: The demonstration of four datasets in general sets.

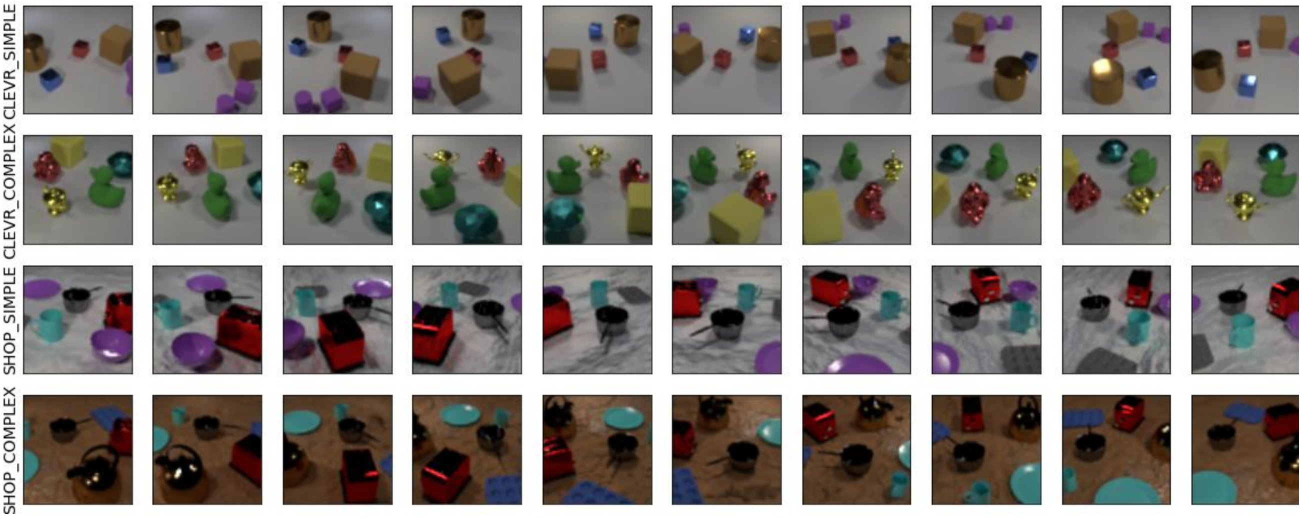


Figure 1: The demonstration of four datasets in test sets.

5 COMPUTATION OF METRICS

In this section, we will introduce all the metrics used in this article, including some matrics not described in the main text. 1) Adjusted Rand Index (ARI) [Hubert and Arabie, 1985] and Adjusted Mutual Information (AMI) [Xuan et al., 2010] assess the quality of segmentation, i.e., how accurately images are partitioned into different objects and background. Previous work usually evaluates ARI and AMI only at pixels belong to objects, and how accurately background is separated from objects is unclear. We evaluate ARI and AMI under two conditions. ARI-A and AMI-A are computed considering both objects and background, while ARI-O and AMI-O are computed considering only objects. 2) Intersection over Union (IoU) and F_1 score (F_1) assess the quality of amodal segmentation, i.e., how accurately complete shapes of objects are estimated. 3) Count assesses the accuracy of the estimated number of objects. 4) Object Ordering Accuracy (OOA) as used in [Yuan et al., 2019] assesses the accuracy of the estimated pairwise ordering of objects. We now desrcibe the mathematical computation in the following.

5.1 DEFINITION

Suppose the test sets have I visual scenes and each visual scene includes T images from different viewpoints, let $\hat{K}_i$ be the be the real maximum number of objects appearing in the ith visual scene (the total number of objects appearing in all visual angles), and let K_i be estimated maximum number of objects appearing in the ith visual scene. note that $\hat{K}_i$ and K_i are not necessarily equal. $\hat{\boldsymbol{r}}_i \in \{0,1\}^{T\times(\hat{K}_i+1)\times N}$ and $\boldsymbol{r}_i \in \{0,1\}^{T\times(\hat{K}_i+1)\times N}$ respectively represent the real and estimated one-hot vector of the T viewpoints in the ith scene corresponding to the pixel-wise partitions (including the foreground and background). $\mathcal{D}_t^i$ denotes the index sets that belong to the object areas in the tth viewpoint of the ith scene, i.e., $\mathcal{D}_t^i = \{n \mid x_{t,n}^i \in \text{object areas}\}$. Let $\hat{U}_{t,k}^i$ be the real index sets w.r.t. object k in the tth viewpoint of the ith scene, i.e., $\hat{U}_{t,k}^i = \{n \mid \boldsymbol{x}_{t,n}^i \in \text{areas of object } k\}$ $(0 \leq k \leq \hat{K}_i)$. Let $U_{t,k}^i$ be the estimated index sets w.r.t. object k in the tth viewpoint of the ith scene. $\hat{U}_{t,k}^i = \{n \mid \hat{\boldsymbol{x}}_{t,n}^i \in \text{areas of object } k\}$ $(0 \leq k \leq \hat{K}_i)$, where $\hat{\boldsymbol{x}}$ is the reconstructed image. Let $\hat{\boldsymbol{s}}_{1:}$

5.2 ADJUSTED RAND INDEX

The computation of Adjusted Rand Index (ARI) is described as:

$$\text{ARI} = \frac{1}{I}\sum_{i=1}^{I}\frac{b_{\text{all}}^i - b_{\text{row}}^i \cdot b_{\text{col}}^i / c^i}{\left(b_{\text{row}}^i + b_{\text{col}}\right)/2 - b_{\text{row}}^i \cdot b_{\text{col}}^i / c^i} \tag{45}$$

In order to explain the meaning of each variable above in detail, $C(x,y)$ is used here to represent the combination number, i.e., $C(x,y) = \frac{x!}{(x-y)!y!}$; $v_{\hat{k},k}^i$ denotes the dot product, i.e., $v_{\hat{k},k}^i = \sum_{(t,n)\in\mathcal{S}}(\hat{r}_{t,k,n} \cdot r_{t,k,n})$, b_{row}^i, b_{col}^i and c^i in Eq 45 are described as:

$$b^i_{\text{all}} = \sum\nolimits_{\hat{k}=0}^{\hat{K}_i} \sum\nolimits_{k=0}^{K} C\Big(v^i_{\hat{k},k}, 2\Big) \tag{46}$$

$$b^i_{\text{row}} = \sum\nolimits_{\hat{k}=0}^{\hat{K}_i} C\Big(\sum\nolimits_{k=0}^{K} v^i_{\hat{k},k}, 2\Big) \tag{47}$$

$$b^i_{\text{col}} = \sum\nolimits_{k=0}^{K} C\Big(\sum\nolimits_{\hat{k}=0}^{\hat{K}_i} v^i_{\hat{k},k}, 2\Big) \tag{48}$$

$$c^i = C\big(\sum\nolimits_{\hat{k}=0}^{\hat{K}_i} \sum\nolimits_{(t,n)\in\mathcal{S}} \hat{r}^i_{t,\hat{k},n}, 2\big) \tag{49}$$

where $\mathcal{S} = \big\{1, 2, ..., T\big\} \times \big\{1, 2, ..., N\big\}$. When computing ARI-O, pixels in $\mathcal{S}$ that do not belong to objects will be removed; When ARI-A is calculated, all pixels in $\mathcal{S}$ will be used.

5.3 ADJUSTED MUTUAL INFORMATION

$$\text{AMI} = \frac{1}{I} \sum_{i=1}^{I} \sum_{t=1}^{T} \frac{\text{MI}(\hat{\boldsymbol{l}}^i, \boldsymbol{l}^i) - \mathbb{E}\big[\text{MI}(\hat{\boldsymbol{l}}^i, \boldsymbol{l}^i)\big]}{\big(\text{H}(\hat{\boldsymbol{l}}^i) + \text{H}(\boldsymbol{l}^i)\big)/2 - \mathbb{E}\big[\text{MI}(\hat{\boldsymbol{l}}^i, \boldsymbol{l}^i)\big]} \tag{50}$$

where $\hat{\boldsymbol{l}}^i \in \mathbb{R}^{T\times(\hat{K}_i+1)}$. $\hat{\boldsymbol{l}}^i_t$ represents the probability distribution of the tth viewpoint in the ith visual scene, i.e.,$\hat{l}^i_t = \big\{|\hat{U}_{t,k}|/|\mathcal{D}^i_t| \mid 0 \leq k \leq \hat{K}_i\big\}$. H and MI respectively represent the entropy and mutual information of the distribution.

$$\text{H}(\hat{\boldsymbol{l}}^i) = -\sum\nolimits_{k=0}^{\hat{K}_i} \sum\nolimits_{k=1}^{T} \hat{l}^i_{t,k} \log \hat{l}^i_{t,k} \tag{51}$$

$$\text{H}(\boldsymbol{l}^i) = -\sum\nolimits_{k=0}^{K_i} \sum\nolimits_{k=1}^{T} l^i_{t,k} \log l^i_{t,k} \tag{52}$$

$$\text{MI}(\hat{\boldsymbol{l}}^i, \boldsymbol{l}^i) = \sum_{m=0}^{\hat{K}_i} \sum_{n=0}^{K_i} \sum_{t=1}^{T} p^i_{t,m,n} \log\Big(\frac{p^i_{t,m,n}}{\hat{l}^i_{t,m} \cdot l^i_{t,n}}\Big) \tag{53}$$

where $\hat{l}^i_{t,k}$ and $l^i_{t,k}$ respectively represent the probability that the pixel in the ith image is partitioned to object k. $p^i_{t,m,n}$ indicates the probability that pixels in the tth frame of the ith image are divided into objects m in the first set and objects n in the second set.$p^i_{t,m,n}$ is calculated as follows:

$$p^i_{t,m,n} = \frac{o^i_{t,m,n}}{|\mathcal{D}^i_t|} = \frac{|\hat{U}^i_{t,m} \cap U^i_{t,n}|}{|\mathcal{D}^i_t|} \tag{54}$$

The matrix $\boldsymbol{o}^i_t \in \mathbb{R}^{(\hat{K}_i+1)\times(K_i+1)}$ is called the contingency table. And the expectation of MI can be analytically computed:

$$\mathbb{E}\big[\text{MI}(\hat{\boldsymbol{l}}^i, \boldsymbol{l}^i)\big] = \sum_{t=1}^{T} \sum_{m=0}^{\hat{K}_i} \sum_{n=0}^{K_i} \sum_{k=(a^i_{t,m}+b^i_{t,n}-N)^+}^{\min(a^i_{t,m}, b^i_{t,n})} \frac{k}{N} \cdot \log\Big(\frac{N\times k}{a^i_{t,m}\times b^i_{t,n}}\Big) \\ \frac{a^i_{t,m}!b^i_{t,n}!(N-a^i_{t,m})!(N-b^i_{t,n})}{N!k!(a^i_{t,m}-k)!(b^i_{t,n}-k)!(N-a^i_{t,m}-b^i_{t,n}+k)!} \tag{55}$$

where $(a^i_{t,m} + b^i_{t,n} - N)^+ = \max(1, a^i_{t,m} + b^i_{t,n} - N)$, $a^i_{t,m}$ and $b^i_{t,n}$ respectively represent the sum of rows and columns w.r.t. $\boldsymbol{o}^i_t$:

$$a^i_{t,m} = \sum\nolimits_{n=0}^{K_i} o^i_{t,m,n}, \quad b^i_{t,n} = \sum\nolimits_{m=0}^{\hat{K}_i} o^i_{t,m,n} \tag{56}$$

5.4 INTERSECTION OVER UNION

In order to compute the Intersection over Union (IoU), we should define two variables: $\hat{\boldsymbol{s}}^i \in \left[0, 1\right]^{T \times \hat{K}_i \times N}$ and $\boldsymbol{s}^i \in \left[0, 1\right]^{T \times K_i \times N}$, IoU from multiple viewpoints requires object index matching under multiple viewpoints, that is,

$$\boldsymbol{\xi}^i = \text{argmax}^i_{\boldsymbol{\xi}^i \in \boldsymbol{\Omega}} \sum\nolimits_{t=1}^{T} \sum\nolimits_{k=1}^{\hat{K}_i} \sum\nolimits_{n=1}^{N} \hat{r}^i_{t,k,n} \cdot r^i_{t,\xi^i_k,n} \tag{57}$$

where $\boldsymbol{\Omega}^i$ is the full arrangement of all object indexes. And the computation of IoU is desceribed as:

$$\text{IoU} = \frac{1}{I} \sum_{i=1}^{I} \frac{1}{\hat{K}_i} \sum_{k=1}^{\hat{K}_i} \frac{\sum_{t=1}^{T} \sum_{n=1}^{N} \min(\hat{s}^i_{t,k,n}, s^i_{t,k,n})}{\sum_{t=1}^{T} \sum_{n=1}^{N} \max(\hat{s}^i_{t,k,n}, s^i_{t,k,n})} \tag{58}$$

5.5 F1 SCORE

F_1 Score is computed as:

$$F_1 = \frac{1}{I} \sum_{i=1}^{I} \frac{1}{\hat{K}_i} \sum_{k=1}^{\hat{K}_i} \frac{2 \cdot \sum_{t=1}^{T} \sum_{n=1}^{N} \min(\hat{s}^i_{t,k,n}, s^i_{t,k,n})}{\sum_{t=1}^{T} \sum_{n=1}^{N} \left(\min(\hat{s}^i_{t,k,n}, s^i_{t,k,n}) + \max(\hat{s}^i_{t,k,n}, s^i_{t,k,n})\right)} \tag{59}$$

5.6 COUNT

$\hat{K}_i$ and K_i represent the real/estimated object numbers in the ith visual scene. When it comes to the model with $\boldsymbol{z}^{\text{pres}}$, K_i is computed through $K_i = \sum_{i=1}^{\hat{K}_i} \text{z}^{\text{pres}}_i$. For the model without $\boldsymbol{z}^{\text{pres}}$, the method to determine the number is: if a layer has no object pixels, the number of objects will not be included. Let δ denotes the Kronecker delta function, and Count is computed as follows:

$$\text{Count} = \frac{1}{I} \sum\nolimits_{i=1}^{I} \delta_{\hat{K}_i, K_i} \tag{60}$$

5.7 OBJECT ORDERING ACCURACY

Another set of vectors should be introduced to compute the ordering relationship of objects. Let $\hat{o}^i_{t,k_1,k_2} \in \left\{0, 1\right\}$, $o^i_{t,k_1,k_2} \in \left\{0, 1\right\}$ respectively represent the real/estimated order of the k_1 object and the k_2 object. Here, the index order of the real object matches the estimated object index one by one. This matching relationship is obtained through the formula Eq 57. The estimated object index will be redirected to $\boldsymbol{\xi}^i$. Because it is difficult to estimate the depth ordering of two objects if they do not overlap, the following OOA calculation measures the importance of different object pairs with different weights:

$$\text{OOA} = \frac{1}{I} \sum_{i=1}^{I} \frac{1}{T} \sum_{t=1}^{T} \frac{\sum_{k_1=1}^{\hat{K}_i - 1} \sum_{k_2=k_1+1}^{\hat{K}_i} w^i_{t,k_1,k_2} \delta_{\hat{o}^i_{t,k_1,k_2}, o^i_{t,k_1,k_2}}}{\sum_{k_1=1}^{\hat{K}_i - 1} \sum_{k_2=k_1+1}^{\hat{K}_i} w^i_{t,k_1,k_2}} \tag{61}$$

The weight of object pairs k_1 and k_2 is calculated as follows:

$$w^i_{m,k_1,k_2} = \sum\nolimits_{n=1}^{N} \hat{s}^i_{t,k,n} \cdot s^i_{t,k,n} \tag{62}$$

The value of w^i_{t,k_1,k_2} reflects the overlapping area of two different object shapes. When the overlapping area is larger, it is easier to do depth sorting, that is, its contribution to the measurement of OOA is greater; On the contrary, when there is little or no overlap ($w^i_{t,k_1,k_2} \to 0$), it has little impact on the measurement of OOA.

6 HYPERPARAMETER CONFIGURATION

We detail the hyperparameter configuration in this section, including the network design, learning rate, temperature schedule, e.t.c. . During the training, the standard deviation σ_x of the likelihood function is chosen to be 0.2. The object slot number K (i.e., the maximum number that may appear in the visual scene) is set to be 7. The dimension of $\boldsymbol{\lambda}$,$\boldsymbol{z}^{\text{view}}$, $\boldsymbol{z}^{\text{obj}}$, $\boldsymbol{z}^{\text{bck}}$ is respectively 5/3/64/16. The fixed standard deviation σ_w of $\boldsymbol{\lambda}$ is 0.8. The hyperparameter α of $\boldsymbol{z}^{\text{view}}$ is 12.6. During the stage 1 training, since there are no labels for the model, it's very difficult to extract the feature from multiple views. We use the warm-up schedule, i.e. single-view training. We used single-view training in the first 30k steps to better initialize the network parameters (single-view learning is easier than multi-view learning), then the model gradually transits to multiple views (for example, you can directly jump to 4 viewpoints, or 2→4). In the sequential extention of Slot Attention [Locatello et al., 2020], we have additional hyperparameters. $\boldsymbol{y}^{\text{view}}$ and $\boldsymbol{y}^{\text{attr}}$ respectively have 8 and 128 dimensions, and D_{key} and D_{val} are 64 and 136. The iteration step is set to be 3. In the learning, the batch size is chosen to be 32. The initial learning rate is $4e-4$, and is decayed exponentially with a factor 0.5 every 50,000 steps. In the first 10,000 training steps, the learning rate is multiplied by a factor that is increased linearly from 0 to 1. For the temperature of $\boldsymbol{z}^{\text{pres}}$, the logarithmic temperature decreases linearly from 10 to 0.5 in the first 150k steps. Now, let us introduce the additional hyperparameters of Stage 1 training and Stage 2 training.

Stage 1 training During the Stage 1 training, since the prior of view latent variables are standard Gaussian distributions, the model does not need to introduce GPs with neural networks. We set the same weight for all KLs. The weight increases linearly from 0 to 1 in the first 100k steps to stabilize the training.

Stage 2 training During the Stage 2 training, we aim to learn the view function of t. Since we use the pretrained model in phase 1, its feature extraction for objects and backgrounds has been stable. At this time, we need to adjust the prior of view latent variables. $\boldsymbol{A}$ in the $\boldsymbol{\lambda}$ prior is implemented by a single full-connected layer without bias. The neural networks of GPs will be detailed in the description of the neural network design. For course learning, we realize it by gradually increasing the view numbers. In the training process, we set different courses for different datasets. In short, for every tens of thousands of steps, two images corresponding to the additional viewpoints will be added as inputs. SHOP has more iterations per course than CLEVR.

Neural Network Design We list the neural networks used in the model:

- f_{feat} denotes the view encoder
  - 3 × 3, stride=(2,2), padding=(1,1), Conv(3,64), ReLU
  - 3 × 3, stride=(2,2), padding=(1,1), Conv(64,64), ReLU
  - 3 × 3, stride=(2,2), padding=(1,1), Conv(64,64)
- the position encoder denotes the 3D postion layer, the xy position and the time t will be encoded to the feature with 192 dimensions, then make the mapping with the conv layer
  - 1 × 1, Conv(192,64)
- View Transformer encoder (before downsample) is the same configuration as SIMONe [Kabra et al., 2021] with 4 layers and 4 heads
- 2 × 2 Downsample
- Spatial Transformer Encoder (after downsample) is set with 4 layers and 4 heads
- f_{λ} denotes the encoder that maps the feature to the mean and variance of $\boldsymbol{\lambda}$
  - Linear(66,64), ReLU
  - Linear(64,32), ReLU
  - Linear(32,15), ReLU
  - Reshape(3,5) (3 corresponds to view, 5 corresponds to the spatial attribues)
- Slot Attention encoder f_{sa}
  - position embedding layer: 1 × 1 Conv(4,64)
  - 5 × 5, padding=(2,2), stride=(1,1), Conv(3,64), ReLU
  - 5 × 5, padding=(2,2), stride=(1,1), Conv(64,64), ReLU
  - 5 × 5, padding=(2,2), stride=(1,1), Conv(64,64), ReLU

- 5 × 5, padding=(2,2), stride=(1,1), Conv(64,64), ReLU
- LayerNorm(64)
- Linear(64,64), ReLU
- Linear(64,64)

- sequential extention of Slot Attention
  - layerNorm(64) (input)
  - layerNorm(136) (query)
  - layerNorm(146) (residual)
  - query: Linear(136,64)
  - key: Linear(64,64)
  - val: Linear(64,136)
  - gru(136,136)
  - residual net: Linear(136,128), ReLU, Linear(128,136)
- view mapping layer that maps the feature extracted from the Transformer to the view slot in the Slot Attention - Linear(64,8)
- f_{view} denotes the view encoder that maps the view slot to the mean and variance of $\boldsymbol{z}^{\text{view}}$
  - Linear(8,512), ReLU
  - Linear(512,512), ReLU
  - Linear(512,6)
- neural networks that correspond to the learnable GP kernal: LargeFeatureExtractor × 3. The design of LargeFeatureExtractor is as follows:
  - Linear(5,32), ReLU
  - Linear(32,32), ReLU
  - Linear(32,64), ReLU
  - Linear(64,64), ReLU
  - Linear(64,8)
- f_{obj} (the output is splited to [128,1,1,1]) denotes the encoder that encode the object slots to the parameters of object latent variables.
  - Linear(128,512), ReLU
  - Linear(512,512), ReLU
  - Linear(512,131)
- f_{bck} denotes the encoder that encode the background slot to the parameters of the background latent variable.
  - Linear(128,512), ReLU
  - Linear(512,512), ReLU
  - Linear(512,32)
- g_{ord} outputs the order value of each object from multiple viewpoints
  - Linear(67,512), ReLU
  - Linear(512,512), ReLU
  - Linear(512,1)
- g_{obj} denotes the object decoder
  - Linear(67,4096), ReLU
  - Linear(4096,4096), ReLU
  - Linear(4096,8192),ReLU
  - Flatten()
  - 2 × Interpolate; 5 × 5, padding=(2,2), stride=(1,1), Conv(128,128); ReLU
  - 5 × 5, padding=(2,2), stride=(1,1), Conv(128,64); ReLU
  - 2 × Interpolate; 5 × 5, padding=(2,2), stride=(1,1), Conv(64,64); ReLU

- 5 × 5, padding=(2,2), stride=(1,1), Conv(64,32); ReLU
- 2 × Interpolate; 5 × 5, padding=(2,2), stride=(1,1), Conv(32,32); ReLU
- 3 × 3, padding=(1,1), Conv(32,4)

- g_{bck} denotes the background decoder
  - Linear(11,512), ReLU
  - Linear(512,512), ReLU
  - Linear(512,256), ReLU
  - Flatten()
  - 4 × Interpolate; 5 × 5, padding=(2,2), stride=(1,1), Conv(16,16); ReLU
  - 5 × 5, padding=(2,2), stride=(1,1), Conv(16,16); ReLU
  - 4 × Interpolate; 5 × 5, padding=(2,2), stride=(1,1), Conv(16,16); ReLU
  - 3 × 3, padding=(1,1), Conv(16,3)

7 ADDITIONAL EXPERIMENTAL RESULTS

In this section, we add more visualization results and comparson results of four datasets, including the observation evaluation and prediction evaluation. Since we use the Stage 1 results to evaluate the quality of the representations, we compare our proposed model with three models called MulMON [Li et al., 2020], SIMONe [Kabra et al., 2021] and OCLOC [Yuan et al., 2022], where MulMON is trained and tested with viewpoint annotations and SIMONe and OCLOC are unsupervised generative We use the Stage 2 results to evaluate the accuracy of novel viewpoints' predictions with the time t. As far as we know, there is no model that can only use time t to predict novel viewpoints. For this reason, we compare it with MulMON model based on viewpoint annotations. We do not compare the proposed model with SIMONe and OCLOC in terms of the prediction since both of them cannot make predictions from novel viewpoints. Note that we compare almost all the metrics, where the computation of IoU and F_1 in MulMON and SIMONe is based on the mask rather than the complete shape, there will be errors to some extent. Nevetheless, we make a complete table of these data.

7.1 UNSUPERVISED LEARNING FROM MULTIPLE VIEWPOINTS

Figures 3, 4, 5 and 6 demonstrate the compared results of four datasets. We can find that our proposed can 1) separate the background from the foreground, which is not reflected in MulMON and SIMONe 2) can completely reconstruct the occluded object from some viewpoints. 3) can effectively remove shadows.This problem is very serious in OCLOC, and we solved it effectively.

The performance of the model is evaluated quantitatively in terms of segmentation, complete shape, occlusion and object counting. Tables 3 and 5 demonstrate the comparison results in 4 views and 8 views, our proposed outperforms the remaining models in multiple aspects. Moreover, we compared the models in the generalization set with more objects and ;larger occlusion rate. Our model is still better than many unsupervised models, and can compete with MulMON with viewpoint annotations. Figure 7 describes the visualized results in general sets, we can find that the model performs well. And Tables 4 and 6 demonstrate the qualitative results of general sets.

7.2 PREDICTION

We have fixed a number of viewpoints to make a fair comparison of prediction performance. Two tested mode called mode 1 and mode 2 are selected. In mode 1, the predicted viewpoints are inserted into the observed viewpoints. And in mode 2, the predicted viewpoints are completely out of the middle. For the two modes, we tested the prediction performance with 6/7/8/9 observed views. Figures 8, 10, 12 and 14 demonstrate the prediciton results of four datasets testing with mode 1. In addition to the prediction of novel viewpoints, our model can also recontrcution additional occlusion completion, which MulMON cannot do. Figures 9, 11, 13 and 15 are tested with mode 2. From the prediciton results, we can see that the farther away from the point of GP function, the worse the reconstruction performance will be.

We evaluate the qualitative results from different observed view numbers. Tables 7, 9, 11, 13, 8, 10, 12 and 14 show the qualitative results on multiple aspects. With more GP points, our model is getting better and better in fitting function. When the observd view number is only 6, our model is slightly worse than MulMON, while when the observed view number becomes more (such as up to 8), our method can make better predictions due to better function fitting, so it is better than MulMON in multiple metrics.

Dataset	Method	ARI-A↑	AMI-A↑	ARI-O↑	AMI-O↑	IoU↑	F1↑	OCA↑	OOA↑
CLEVR-SIMPLE	MulMON	0.658±1e-3	0.603±1e-3	0.969±1e-3	0.956±1e-3	0.615±4e-3	0.741±4e-3	0.606±4e-2	N/A
	SIMONe	0.086±5e-5	0.313±9e-5	0.947±1e-4	0.924±2e-4	0.449±1e-4	0.601±2e-4	0.000±0e-0	N/A
	OCLOC	0.541±2e-3	0.512±2e-3	0.935±5e-3	0.930±4e-3	0.475±4e-3	0.629±4e-3	0.532±3e-2	0.955±1e-2
	Ours	**0.830±3e-3**	**0.736±3e-3**	**0.973±7e-3**	**0.968±4e-3**	**0.656±3e-3**	**0.781± 4e-3**	**0.704±3e-2**	**0.968±1e-2**
CLEVR-COMPLEX	MulMON	0.552±9e-3	0.533±4e-3	0.941±3e-3	0.923±2e-3	0.554±3e-3	0.689±4e-3	0.612±3e-2	N/A
	SIMONe	0.073±3e-5	0.299±8e-5	0.939±2e-4	0.912±3e-4	0.396±5e-5	0.547±6e-5	0.000±0e-0	N/A
	OCLOC	0.396±1e-3	0.419±1e-3	0.915±4e-3	0.905±4e-3	0.375±3e-3	0.523±3e-3	0.676±2e-2	0.917±1e-2
	Ours	**0.759±2e-3**	**0.657±3e-3**	**0.963±4e-3**	**0.959±3e-3**	**0.569±6e-3**	**0.708±7e-3**	**0.694±2e-2**	**0.952±1e-2**
SHOP-SIMPLE	MulMON	0.435±2e-2	0.539±8e-3	0.894±5e-3	0.878±2e-3	0.596±9e-3	0.725±9e-3	0.148±4e-2	N/A
	SIMONe	0.201±2e-4	0.437±2e-4	0.757±1e-4	0.805±1e-4	0.488±7e-5	0.633±7e-5	0.000±0e-0	N/A
	OCLOC	0.650±4e-3	0.607±4e-3	0.918±6e-3	0.910±4e-3	0.609±4e-3	0.737±5e-3	0.448±5e-2	0.695±2e-2
	Ours	**0.816±2e-3**	**0.739±2e-3**	**0.957±2e-3**	**0.954±1e-3**	**0.668±3e-3**	**0.780±3e-3**	**0.528±8e-2**	**0.790±2e-2**
SHOP-COMPLEX	MulMON	0.599±2e-2	0.595±6e-3	0.872±4e-3	0.863±2e-3	0.630±4e-3	0.751±4e-3	0.314±4e-2	N/A
	SIMONe	0.185±6e-5	0.443±8e-5	0.796±7e-5	0.840±9e-5	0.535±1e-4	0.675±1e-4	0.000±0e-0	N/A
	OCLOC	0.342±1e-3	0.305±1e-3	0.380±5e-3	0.495±4e-3	0.249±3e-3	0.360±4e-3	0.160±4e-2	0.624±2e-2
	Ours	**0.796±4e-3**	**0.714±3e-3**	**0.946±7e-3**	**0.941±4e-3**	**0.654±4e-3**	**0.771±4e-3**	**0.518±2e-2**	**0.852±8e-3**

Table 3: The comparison results of multiple aspects on test sets (training on 4 views and testing on 4 views). All test values are evaluated 5 times, recorded with mean and standard deviation.

Dataset	Method	ARI-A↑	AMI-A↑	ARI-O↑	AMI-O↑	IoU↑	F1↑	OCA↑	OOA↑
CLEVR-SIMPLE	MulMON	0.584±8e-4	0.606±9e-4	**0.939±2e-3**	**0.933±1e-3**	0.542±3e-3	0.671±4e-3	**0.440±5e-2**	N/A
	SIMONe	0.111±6e-5	0.409±1e-4	0.912±3e-4	0.885±3e-4	0.430±8e-5	0.573±8e-5	0.000±0e-0	N/A
	OCLOC	0.406±2e-3	0.489±3e-3	0.863±6e-3	0.872±4e-3	0.397±5e-3	0.541±7e-3	0.250±3e-2	0.897±8e-3
	Ours	**0.763±8e-4**	**0.706±1e-3**	0.931±2e-3	0.931±1e-3	**0.569±3e-3**	**0.691±3e-3**	0.390±6e-2	**0.936±8e-3**
CLEVR-COMPLEX	MulMON	0.477±3e-3	0.539±7e-4	0.906±2e-3	0.897±1e-3	0.469±1e-3	0.601±2e-3	0.326±4e-2	N/A
	SIMONe	0.090±3e-5	0.392±6e-5	0.914±2e-4	0.887±2e-4	0.387±5e-5	0.528±5e-5	0.000±0e-0	N/A
	OCLOC	0.187±9e-4	0.388±1e-3	0.829±6e-3	0.845±3e-3	0.290±8e-4	0.424±1e-3	0.316±2e-2	0.853±4e-3
	Ours	**0.676±2e-3**	**0.630±3e-3**	**0.917±6e-3**	**0.919±4e-3**	**0.496±6e-3**	**0.628±7e-3**	**0.390±4e-2**	**0.917±8e-3**
SHOP-SIMPLE	MulMON	0.509±9e-3	0.590±3e-3	0.871±3e-3	0.873±1e-3	0.565±4e-3	**0.694±5e-3**	0.316±5e-2	N/A
	SIMONe	0.200±9e-5	0.454±8e-5	0.709±1e-4	0.763±1e-4	0.396±2e-4	0.527±2e-4	0.000±0e-0	N/A
	OCLOC	0.459±3e-3	0.525±3e-3	0.817±6e-3	0.838±4e-3	0.481±6e-3	0.612±7e-3	0.146±2e-2	0.636±2e-2
	Ours	**0.737±2e-3**	**0.696±2e-3**	**0.921±4e-3**	**0.920±2e-3**	**0.570±5e-3**	0.689±6e-3	**0.336±3e-2**	**0.816±1e-2**
SHOP-COMPLEX	MulMON	0.563±6e-3	0.594±2e-3	0.841±7e-3	0.850±3e-3	**0.553±3e-3**	**0.677±3e-3**	0.318±3e-2	N/A
	SIMONe	0.196±3e-5	0.481±7e-5	0.785±1e-4	0.818±2e-4	0.481±1e-4	0.610±1e-4	0.004±5e-3	N/A
	OCLOC	0.230±3e-3	0.277±2e-3	0.301±2e-3	0.453±8e-4	0.179±1e-3	0.269±2e-3	0.172±1e-2	0.557±2e-2
	Ours	**0.706±3e-3**	**0.666±3e-3**	**0.893±3e-3**	**0.893±3e-3**	0.550±6e-3	0.670±6e-3	**0.326±5e-2**	**0.808±6e-3**

Table 4: The comparison results of multiple aspects on general sets (training on 4 views and testing on 4 views). All test values are evaluated 5 times, recorded with mean and standard deviation.

Dataset	Method	ARI-A↑	AMI-A↑	ARI-O↑	AMI-O↑	IoU↑	F1↑	OCA↑	OOA↑
CLEVR-SIMPLE	MulMON	0.632±1e-3	0.582±1e-3	**0.964±9e-4**	0.949±7e-4	**0.596±2e-3**	0.727±3e-3	0.564±2e-2	N/A
	SIMONe	0.106±4e-5	0.310±3e-5	0.910±2e-4	0.887±2e-4	0.398±6e-5	0.555±6e-5	0.000±0e-0	N/A
	OCLOC	0.520±9e-4	0.492±1e-3	0.927±8e-3	0.917±4e-3	0.456±2e-3	0.615±3e-3	0.628±4e-2	0.936±1e-2
	Ours	**0.772±2e-3**	**0.671±2e-3**	0.959±3e-3	**0.954±3e-3**	0.595±5e-3	**0.733±5e-3**	**0.594±5e-2**	**0.953±1e-2**
CLEVR-COMPLEX	MulMON	0.521±1e-2	0.509±6e-3	0.929±2e-3	0.908±2e-3	**0.534±4e-3**	**0.672±4e-3**	**0.604±3e-2**	N/A
	SIMONe	0.092±1e-5	0.316±3e-5	0.914±3e-4	0.878±3e-4	0.423±2e-5	0.575±2e-5	0.000±0e-0	N/A
	OCLOC	0.366±1e-3	0.375±1e-3	0.827±8e-3	0.824±3e-3	0.351±2e-3	0.500±3e-3	0.168±5e-2	0.891±1e-2
	Ours	**0.696±2e-3**	**0.592±2e-3**	**0.941±3e-3**	**0.932±3e-3**	0.509±4e-3	0.657±4e-3	0.550±8e-2	**0.930±8e-3**
SHOP-SIMPLE	MulMON	0.435±1e-2	0.530±5e-3	0.883±6e-3	0.863±4e-3	0.587±4e-3	0.719±4e-3	0.160±4e-2	N/A
	SIMONe	0.135±7e-5	0.321±1e-4	0.553±1e-4	0.581±2e-4	0.330±5e-5	0.462±6e-5	0.000±0e-0	N/A
	OCLOC	0.663±3e-3	0.609±3e-3	0.913±4e-3	0.897±3e-3	0.619±6e-3	0.746±7e-3	0.388±2e-2	0.728±1e-2
	Ours	**0.803±7e-4**	**0.726±6e-4**	**0.958±1e-3**	**0.954±1e-3**	**0.656±6e-4**	**0.774±7e-4**	**0.528±5e-2**	**0.789±4e-3**
SHOP-COMPLEX	MulMON	0.585±1e-2	0.583±4e-3	0.871±2e-3	0.859±2e-3	0.625±5e-3	0.750±5e-3	0.330±5e-2	N/A
	SIMONe	0.106±5e-5	0.234±2e-5	0.335±3e-4	0.388±2e-4	0.216±1e-4	0.329±2e-4	0.000±0e-0	N/A
	OCLOC	0.350±3e-3	0.273±2e-3	0.293±5e-3	0.408±6e-3	0.215±3e-3	0.321±4e-3	0.064±5e-3	0.579±1e-2
	Ours	**0.786±4e-3**	**0.703±3e-3**	**0.949±4e-3**	**0.940±3e-3**	**0.662±6e-3**	**0.781±6e-3**	**0.434±3e-2**	**0.818±1e-2**

Table 5: The comparison results of multiple aspects on test sets (training on 8 views and testing on 8 views). All test values are evaluated 5 times, recorded with mean and standard deviation.

Dataset	Method	ARI-A↑	AMI-A↑	ARI-O↑	AMI-O↑	IoU↑	F1↑	OCA↑	OOA↑
CLEVR-SIMPLE	MulMON	0.554±1e-3	0.584±2e-3	**0.924±3e-3**	**0.920±3e-3**	**0.522±5e-3**	**0.656±6e-3**	**0.376±3e-2**	N/A
	SIMONe	0.132±4e-5	0.368±4e-5	0.787±7e-5	0.777±1e-4	0.326±3e-5	0.452±3e-5	0.020±1e-2	N/A
	OCLOC	0.393±2e-3	0.473±2e-3	0.844±6e-3	0.853±4e-3	0.386±4e-3	0.533±5e-3	0.254±3e-2	0.891±1e-2
	Ours	**0.692±3e-3**	**0.635±4e-3**	0.890±1e-2	0.897±6e-3	0.505±7e-3	0.637±8e-3	0.366±4e-2	**0.913±1e-2**
CLEVR-COMPLEX	MulMON	0.458±4e-3	0.523±8e-4	**0.893±2e-3**	**0.882±1e-3**	**0.459±1e-3**	**0.595±2e-3**	0.322±7e-2	N/A
	SIMONe	0.109±1e-5	0.375±3e-5	0.814±6e-5	0.795±9e-5	0.368±7e-5	0.500±1e-4	0.000±0e-0	N/A
	OCLOC	0.173±2e-3	0.349±1e-3	0.730±5e-3	0.761±2e-3	0.267±2e-3	0.399±3e-3	0.084±3e-2	0.827±1e-2
	Ours	**0.597±9e-4**	**0.560±6e-4**	0.865±4e-3	0.873±3e-3	0.428±3e-3	0.569±5e-3	**0.352±3e-2**	**0.874±4e-3**
SHOP-SIMPLE	MulMON	0.490±5e-3	0.575±1e-3	0.865±3e-3	0.859±1e-3	0.557±2e-3	**0.688±3e-3**	**0.360±5e-2**	N/A
	SIMONe	0.105±5e-5	0.276±5e-5	0.418±1e-4	0.467±1e-4	0.193±5e-5	0.290±9e-5	0.002±4e-3	N/A
	OCLOC	0.465±3e-3	0.513±3e-3	0.798±3e-3	0.812±3e-3	0.475±4e-3	0.605±4e-3	0.128±4e-2	0.614±5e-3
	Ours	**0.723±2e-3**	**0.683±2e-3**	**0.914±2e-3**	**0.915±1e-3**	**0.561±3e-3**	0.685±3e-3	0.310±5e-2	**0.827±1e-2**
SHOP-COMPLEX	MulMON	0.561±7e-3	0.588±3e-3	0.839±3e-3	0.844±2e-3	0.560±4e-3	0.689±5e-3	**0.374±5e-2**	N/A
	SIMONe	0.086±2e-5	0.203±4e-5	0.255±1e-4	0.323±9e-5	0.132±3e-5	0.213±6e-5	0.002±4e-3	N/A
	OCLOC	0.246±2e-3	0.243±2e-3	0.213±4e-3	0.371±3e-3	0.158±7e-4	0.246±1e-3	0.096±3e-2	0.570±2e-2
	Ours	**0.702±4e-3**	**0.660±3e-3**	**0.889±3e-3**	**0.889±3e-3**	**0.568±5e-3**	**0.690±6e-3**	0.284±5e-2	**0.823±1e-2**

Table 6: The comparison results of multiple aspects on general sets (training on 8 views and testing on 8 views). All test values are evaluated 5 times, recorded with mean and standard deviation.

Dataset	Query	Method	ARI-A↑	AMI-A↑	ARI-O↑	AMI-O↑	IoU↑	F1↑	MSE↓	OOA↑
CLEVR-SIMPLE	1	MulMON	0.667±1e-3	0.606±1e-3	**0.955±2e-3**	0.948±2e-3	**0.619±2e-3**	**0.741±3e-3**	**0.0017±3e-5**	N/A
		Ours	**0.789±8e-3**	**0.699±8e-3**	0.948±8e-3	**0.949±6e-3**	0.593±5e-3	0.720±5e-3	0.0020±1e-4	0.982±1e-2
	2	MulMON	0.632±1e-3	0.581±1e-3	**0.959±2e-3**	**0.947±1e-3**	**0.592±2e-3**	**0.719±3e-3**	**0.0016±3e-5**	N/A
		Ours	**0.760±4e-3**	**0.662±5e-3**	0.946±5e-3	0.945±4e-3	0.572±3e-3	0.711±4e-3	0.0020±6e-5	0.973±9e-3
	4	MulMON	0.618±9e-4	0.569±9e-4	**0.955±1e-3**	**0.940±1e-3**	**0.580±2e-3**	**0.710±3e-3**	**0.0016±3e-5**	N/A
		Ours	**0.753±4e-3**	**0.648±4e-3**	0.942±6e-3	0.937±4e-3	0.563±4e-3	0.706±5e-3	0.0020±9e-5	0.964±8e-3
CLEVR-COMPLEX	1	MulMON	0.564±1e-2	0.539±6e-3	0.941±4e-3	0.935±2e-3	**0.549±5e-3**	**0.678±6e-3**	**0.0019±5e-5**	N/A
		Ours	**0.764±1e-2**	**0.668±1e-2**	**0.961±5e-3**	**0.959±4e-3**	0.528±6e-3	0.666±6e-3	**0.0019±2e-4**	0.973±2e-2
	2	MulMON	0.522±1e-2	0.506±5e-3	0.926±3e-3	0.910±2e-3	**0.526±4e-3**	**0.661±5e-3**	**0.0021±3e-5**	N/A
		Ours	**0.725±7e-3**	**0.625±7e-3**	**0.956±5e-3**	**0.952±5e-3**	0.513±4e-3	0.660±3e-3	**0.0021±1e-4**	0.941±2e-2
	4	MulMON	0.519±1e-2	0.500±5e-3	0.917±4e-3	0.896±3e-3	**0.521±5e-3**	**0.658±5e-3**	**0.0022±3e-5**	N/A
		Ours	**0.721±1e-3**	**0.614±1e-3**	**0.949±4e-3**	**0.940±4e-3**	0.511±2e-3	0.660±3e-3	0.0023±5e-5	0.942±1e-2
SHOP-SIMPLE	1	MulMON	0.457±1e-2	0.528±5e-3	0.867±3e-3	0.862±3e-3	0.555±4e-3	0.675±4e-3	0.0050±6e-5	N/A
		Ours	**0.785±1e-2**	**0.715±9e-3**	**0.950±8e-3**	**0.953±7e-3**	**0.627±8e-3**	**0.739±9e-3**	**0.0035±4e-4**	0.767±2e-2
	2	MulMON	0.425±1e-2	0.515±4e-3	0.868±3e-3	0.850±1e-3	0.545±4e-3	0.668±4e-3	0.0053±8e-5	N/A
		Ours	**0.768±9e-3**	**0.693±9e-3**	**0.948±8e-3**	**0.948±7e-3**	**0.620±8e-3**	**0.743±9e-3**	**0.0037±3e-4**	0.741±4e-2
	4	MulMON	0.410±1e-2	0.507±5e-3	0.864±3e-3	0.837±1e-3	0.544±4e-3	0.667±4e-3	0.0055±8e-5	N/A
		Ours	**0.758±7e-3**	**0.679±7e-3**	**0.943±5e-3**	**0.940±5e-3**	**0.614±7e-3**	**0.739±8e-3**	**0.0040±3e-4**	0.715±2e-2
SHOP-COMPLEX	1	MulMON	0.612±1e-2	0.577±5e-3	0.835±4e-3	0.838±3e-3	0.571±2e-3	0.674±3e-3	0.0043±9e-5	N/A
		Ours	**0.732±1e-2**	**0.663±9e-3**	**0.918±5e-3**	**0.923±3e-3**	**0.580±8e-3**	**0.698±9e-3**	**0.0038±3e-4**	0.835±2e-2
	2	MulMON	0.607±1e-2	0.569±5e-3	0.836±4e-3	0.831±3e-3	0.558±4e-3	0.666±4e-3	0.0046±8e-5	N/A
		Ours	**0.734±1e-2**	**0.656±1e-2**	**0.920±1e-2**	**0.921±6e-3**	**0.583±9e-3**	**0.711±8e-3**	**0.0038±2e-4**	0.796±2e-2
	4	MulMON	0.601±1e-2	0.561±5e-3	0.824±5e-3	0.819±4e-3	0.554±4e-3	0.663±5e-3	0.0049±9e-5	N/A
		Ours	**0.732±7e-3**	**0.646±7e-3**	**0.911±6e-3**	**0.907±5e-3**	**0.579±6e-3**	**0.709±6e-3**	**0.0040±2e-4**	0.749±2e-2

Table 7: The comparison results of prediction on test sets (the test mode is 1, the observed views are 6, and query views are 1, 2, 4). All test values are evaluated 5 times, recorded with mean and standard deviation.

Dataset	Query	Method	ARI-A↑	AMI-A↑	ARI-O↑	AMI-O↑	IoU↑	F1↑	MSE↓	OOA↑
CLEVR-SIMPLE	1	MulMON	0.604±6e-4	0.564±3e-4	**0.962±2e-3**	**0.957±1e-3**	**0.579±3e-3**	**0.710±4e-3**	**0.0013±2e-5**	N/A
		Ours	**0.725±8e-3**	**0.631±6e-3**	0.934±5e-3	0.938±4e-3	0.544±5e-3	0.679±5e-3	0.0022±1e-4	0.958±1e-2
	2	MulMON	0.621±7e-4	0.576±8e-4	**0.960±3e-3**	**0.951±2e-3**	**0.588±3e-3**	**0.718±4e-3**	**0.0015±2e-5**	N/A
		Ours	**0.733±8e-3**	**0.630±8e-3**	0.924±8e-3	0.925±7e-3	0.554±5e-3	0.692±6e-3	0.0024±1e-4	0.958±1e-2
	4	MulMON	0.667±9e-4	0.609±1e-3	**0.965±3e-3**	**0.953±2e-3**	**0.617±4e-3**	**0.742±4e-3**	**0.0014±2e-5**	N/A
		Ours	**0.774±9e-3**	**0.662±9e-3**	0.923±6e-3	0.918±6e-3	0.586±9e-3	0.723±8e-3	0.0022±2e-4	0.963±5e-3
CLEVR-COMPLEX	1	MulMON	0.503±6e-3	0.497±2e-3	0.931±6e-3	0.929±3e-3	**0.524±8e-4**	**0.664±1e-3**	**0.0019±1e-5**	N/A
		Ours	**0.694±9e-3**	**0.601±8e-3**	**0.949±7e-3**	**0.951±5e-3**	0.495±8e-3	0.640±8e-3	0.0025±1e-4	0.932±4e-2
	2	MulMON	0.520±7e-3	0.509±2e-3	**0.930±7e-3**	**0.923±4e-3**	**0.531±1e-3**	**0.670±2e-3**	**0.0020±2e-5**	N/A
		Ours	**0.681±2e-2**	**0.573±2e-2**	0.886±3e-2	0.890±2e-2	0.473±2e-2	0.618±2e-2	0.0032±3e-4	0.879±3e-2
	4	MulMON	0.562±7e-3	0.538±2e-3	**0.934±7e-3**	**0.920±3e-3**	**0.558±9e-4**	**0.694±1e-3**	**0.0020±1e-5**	N/A
		Ours	**0.716±2e-2**	**0.600±2e-2**	0.901±2e-2	0.897±2e-2	0.502±1e-2	0.648±1e-2	0.0031±2e-4	0.910±3e-2
SHOP-SIMPLE	1	MulMON	0.435±1e-2	0.519±5e-3	0.864±9e-3	0.851±5e-3	0.563±2e-3	0.692±3e-3	0.0047±3e-5	N/A
		Ours	**0.749±7e-3**	**0.676±6e-3**	**0.953±7e-3**	**0.954±5e-3**	**0.589±8e-3**	**0.710±9e-3**	**0.0038±2e-4**	0.846±5e-2
	2	MulMON	0.430±1e-2	0.521±5e-3	0.867±7e-3	0.845±3e-3	0.564±2e-3	0.691±3e-3	0.0048±4e-5	N/A
		Ours	**0.746±5e-3**	**0.668±6e-3**	**0.939±1e-2**	**0.938±8e-3**	**0.584±7e-3**	**0.706±8e-3**	**0.0042±3e-4**	0.814±2e-2
	4	MulMON	0.455±1e-2	0.545±5e-3	0.866±9e-3	0.846±3e-3	0.597±2e-3	0.722±3e-3	0.0049±7e-5	N/A
		Ours	**0.784±4e-3**	**0.701±5e-3**	**0.944±5e-3**	**0.939±5e-3**	**0.633±7e-3**	**0.755±7e-3**	**0.0041±2e-4**	**0.819±3e-2**
SHOP-COMPLEX	1	MulMON	0.666±5e-3	0.619±2e-3	0.864±6e-3	0.859±5e-3	**0.631±2e-3**	**0.746±2e-3**	**0.0036±3e-5**	N/A
		Ours	**0.733±1e-2**	**0.658±1e-2**	**0.931±7e-3**	**0.933±5e-3**	0.573±8e-3	0.696±8e-3	0.0037±3e-4	0.827±4e-2
	2	MulMON	0.669±5e-3	0.620±2e-3	0.870±6e-3	0.857±3e-3	**0.631±2e-3**	**0.745±3e-3**	**0.0036±2e-5**	N/A
		Ours	**0.731±1e-2**	**0.648±1e-2**	**0.911±1e-2**	**0.911±7e-3**	0.561±9e-3	0.683±9e-3	0.0040±2e-4	0.773±2e-2
	4	MulMON	0.702±5e-3	0.645±2e-3	0.866±7e-3	0.856±4e-3	**0.663±2e-3**	**0.775±3e-3**	**0.0037±1e-5**	N/A
		Ours	**0.756±1e-2**	**0.668±1e-2**	**0.910±1e-2**	**0.908±9e-3**	0.604±1e-2	0.729±9e-3	0.0041±2e-4	0.791±1e-2

Table 8: The comparison results of prediction on test sets (the test mode is 2, the observed views are 6, and query views are 1, 2, 4). All test values are evaluated 5 times, recorded with mean and standard deviation.

Dataset	Query	Method	ARI-A↑	AMI-A↑	ARI-O↑	AMI-O↑	IoU↑	F1↑	MSE↓	OOA↑
CLEVR-SIMPLE	1	MulMON	0.668±2e-3	0.607±2e-3	0.955±3e-3	0.947±2e-3	**0.619±3e-3**	**0.741±4e-3**	**0.0017±3e-5**	N/A
		Ours	**0.805±1e-2**	**0.715±1e-2**	**0.964±9e-3**	**0.961±7e-3**	0.606±1e-2	0.734±2e-2	0.0018±3e-4	0.977±2e-2
	2	MulMON	0.629±2e-3	0.575±1e-3	0.950±3e-3	0.939±2e-3	**0.586±4e-3**	0.715±5e-3	**0.0016±4e-5**	N/A
		Ours	**0.768±1e-2**	**0.669±1e-2**	**0.958±1e-2**	**0.954±1e-2**	0.580±1e-2	**0.719±1e-2**	0.0018±2e-4	0.979±1e-2
CLEVR-COMPLEX	1	MulMON	0.567±7e-3	0.542±3e-3	0.944±2e-3	0.937±2e-3	**0.554±4e-3**	**0.684±4e-3**	**0.0018±2e-5**	N/A
		Ours	**0.765±9e-3**	**0.669±8e-3**	**0.959±1e-2**	**0.959±9e-3**	0.524±1e-2	0.661±1e-2	0.0019±2e-4	0.950±9e-3
	2	MulMON	0.526±8e-3	0.506±3e-3	0.925±2e-3	0.907±1e-3	**0.528±3e-3**	**0.664±4e-3**	**0.0021±2e-5**	N/A
		Ours	**0.720±6e-3**	**0.619±6e-3**	**0.948±1e-2**	**0.947±8e-3**	0.505±7e-3	0.651±8e-3	0.0023±8e-5	0.915±2e-2
SHOP-SIMPLE	1	MulMON	0.446±8e-3	0.525±4e-3	0.872±5e-3	0.867±3e-3	0.557±6e-3	0.678±6e-3	0.0049±5e-5	N/A
		Ours	**0.796±4e-3**	**0.725±4e-3**	**0.959±4e-3**	**0.961±3e-3**	**0.635±6e-3**	**0.747±6e-3**	**0.0031±8e-5**	0.788±2e-2
	2	MulMON	0.406±9e-3	0.508±5e-3	0.868±7e-3	0.851±4e-3	0.543±5e-3	0.668±5e-3	0.0056±6e-5	N/A
		Ours	**0.776±4e-3**	**0.700±3e-3**	**0.958±4e-3**	**0.957±2e-3**	**0.624±6e-3**	**0.747±5e-3**	**0.0038±1e-4**	0.820±1e-2
SHOP-COMPLEX	1	MulMON	0.619±9e-3	0.587±4e-3	0.849±4e-3	0.851±1e-3	0.590±3e-3	0.696±3e-3	0.0039±4e-5	N/A
		Ours	**0.750±1e-2**	**0.680±1e-2**	**0.923±6e-3**	**0.930±6e-3**	**0.594±1e-2**	**0.711±1e-2**	**0.0033±2e-4**	0.852±1e-2
	2	MulMON	0.613±8e-3	0.578±3e-3	0.843±4e-3	0.841±1e-3	0.580±4e-3	0.691±4e-3	0.0044±7e-5	N/A
		Ours	**0.745±7e-3**	**0.668±6e-3**	**0.923±8e-3**	**0.926±7e-3**	**0.593±1e-2**	**0.719±1e-2**	**0.0036±1e-4**	0.844±1e-2

Table 9: The comparison results of prediction on test sets (the test mode is 1, the observed views are 7, and query views are 1, 2). All test values are evaluated 5 times, recorded with mean and standard deviation.

Dataset	Query	Method	ARI-A↑	AMI-A↑	ARI-O↑	AMI-O↑	IoU↑	F1↑	MSE↓	OOA↑
CLEVR-SIMPLE	1	MulMON	0.640±3e-3	0.594±1e-3	**0.960±3e-3**	0.954±2e-3	**0.606±3e-3**	**0.733±3e-3**	**0.0016±3e-5**	N/A
		Ours	**0.774±8e-3**	**0.683±7e-3**	0.958±5e-3	**0.959±5e-3**	0.600±6e-3	0.731±7e-3	0.0021±1e-4	0.960±1e-2
	2	MulMON	0.661±2e-3	0.609±2e-3	**0.964±3e-3**	**0.955±2e-3**	**0.614±3e-3**	0.739±3e-3	**0.0015±2e-5**	N/A
		Ours	**0.793±6e-3**	**0.694±6e-3**	0.954±6e-3	0.953±5e-3	0.613±5e-3	**0.744±5e-3**	0.0019±1e-4	0.972±8e-3
CLEVR-COMPLEX	1	MulMON	0.543±2e-3	0.528±1e-3	0.931±3e-3	0.929±2e-3	**0.542±3e-3**	**0.678±4e-3**	**0.0021±2e-5**	N/A
		Ours	**0.722±2e-2**	**0.628±2e-2**	**0.934±1e-2**	**0.935±1e-2**	0.511±2e-2	0.651±2e-2	0.0030±4e-4	0.956±2e-2
	2	MulMON	0.563±3e-3	0.543±2e-3	0.935±4e-3	0.927±2e-3	**0.559±3e-3**	**0.694±4e-3**	**0.0021±2e-5**	N/A
		Ours	**0.743±2e-2**	**0.639±1e-2**	**0.936±2e-2**	**0.934±1e-2**	0.527±1e-2	0.669±2e-2	0.0028±2e-4	0.938±2e-2
SHOP-SIMPLE	1	MulMON	0.445±1e-2	0.536±7e-3	0.886±5e-3	0.870±3e-3	0.580±4e-3	0.705±3e-3	0.0048±5e-5	N/A
		Ours	**0.755±1e-2**	**0.686±8e-3**	**0.954±6e-3**	**0.954±4e-3**	**0.608±6e-3**	**0.724±5e-3**	**0.0043±6e-4**	0.766±1e-2
	2	MulMON	0.448±1e-2	0.545±6e-3	0.881±4e-3	0.864±2e-3	0.591±3e-3	0.716±2e-3	0.0048±6e-5	N/A
		Ours	**0.775±8e-3**	**0.700±8e-3**	**0.951±8e-3**	**0.948±7e-3**	**0.629±3e-3**	**0.749±3e-3**	**0.0041±3e-4**	0.816±3e-2
SHOP-COMPLEX	1	MulMON	0.668±5e-3	0.626±2e-3	0.880±5e-3	0.871±4e-3	**0.639±4e-3**	**0.750±4e-3**	**0.0036±3e-5**	N/A
		Ours	**0.745±1e-2**	**0.672±1e-2**	**0.925±8e-3**	**0.928±6e-3**	0.584±8e-3	0.704±9e-3	0.0038±2e-4	0.730±3e-2
	2	MulMON	0.688±6e-3	0.641±3e-3	0.878±6e-3	0.871±4e-3	**0.653±4e-3**	**0.765±5e-3**	**0.0037±5e-5**	N/A
		Ours	**0.765±1e-2**	**0.683±1e-2**	**0.926±1e-2**	**0.926±1e-2**	0.607±9e-3	0.731±8e-3	0.0038±3e-4	0.764±2e-2

Table 10: The comparison results of prediction on test sets (the test mode is 2, the observed views are 7, and query views are 1, 2). All test values are evaluated 5 times, recorded with mean and standard deviation.

Dataset	Query	Method	ARI-A↑	AMI-A↑	ARI-O↑	AMI-O↑	IoU↑	F1↑	MSE↓	OOA↑
CLEVR-SIMPLE	1	MulMON	0.624±5e-4	0.581±6e-4	**0.969±1e-3**	**0.961±1e-3**	**0.594±2e-3**	**0.722±3e-3**	**0.0014±2e-5**	N/A
		Ours	**0.756±1e-2**	**0.669±9e-3**	0.961±7e-3	**0.961±5e-3**	0.570±8e-3	0.705±8e-3	0.0018±2e-4	0.957±2e-2
	2	MulMON	0.600±3e-4	0.559±6e-4	**0.962±1e-3**	0.950±2e-3	**0.577±2e-3**	**0.711±3e-3**	**0.0014±2e-5**	N/A
		Ours	**0.746±6e-3**	**0.650±7e-3**	0.955±5e-3	**0.953±4e-3**	0.561±4e-3	0.703±4e-3	0.0018±9e-5	0.955±9e-3
CLEVR-COMPLEX	1	MulMON	0.513±6e-3	0.513±3e-3	0.934±3e-3	0.931±2e-3	**0.534±3e-3**	**0.669±3e-3**	**0.0019±3e-5**	N/A
		Ours	**0.724±1e-2**	**0.633±9e-3**	**0.959±8e-3**	**0.957±8e-3**	0.521±9e-3	0.666±1e-2	0.0020±2e-4	0.943±2e-2
	2	MulMON	0.501±6e-3	0.494±3e-3	0.915±3e-3	0.902±2e-3	**0.517±3e-3**	**0.656±3e-3**	**0.0020±2e-5**	N/A
		Ours	**0.710±9e-3**	**0.611±8e-3**	**0.955±9e-3**	**0.948±8e-3**	0.506±7e-3	0.655±8e-3	0.0021±2e-4	0.938±2e-2
SHOP-SIMPLE	1	MulMON	0.442±2e-2	0.526±7e-3	0.890±3e-3	0.877±3e-3	0.562±7e-3	0.690±7e-3	0.0046±1e-4	N/A
		Ours	**0.767±2e-3**	**0.699±3e-3**	**0.960±1e-3**	**0.960±2e-3**	**0.605±4e-3**	**0.723±5e-3**	**0.0032±9e-5**	0.856±3e-2
	2	MulMON	0.415±2e-2	0.519±7e-3	0.883±5e-3	0.864±4e-3	0.570±7e-3	0.701±6e-3	0.0049±8e-5	N/A
		Ours	**0.766±5e-3**	**0.693±5e-3**	**0.960±2e-3**	**0.957±2e-3**	**0.620±7e-3**	**0.744±8e-3**	**0.0034±1e-4**	0.737±1e-2
SHOP-COMPLEX	1	MulMON	0.633±1e-2	0.600±6e-3	0.876±6e-3	0.872±3e-3	**0.606±6e-3**	**0.721±5e-3**	0.0038±9e-5	N/A
		Ours	**0.737±9e-3**	**0.666±8e-3**	**0.932±6e-3**	**0.934±4e-3**	0.572±4e-3	0.694±4e-3	**0.0034±2e-4**	0.855±3e-2
	2	MulMON	0.636±1e-2	0.598±5e-3	0.869±7e-3	0.861±3e-3	**0.613±6e-3**	**0.731±5e-3**	0.0040±6e-5	N/A
		Ours	**0.742±6e-3**	**0.663±6e-3**	**0.929±4e-3**	**0.928±3e-3**	0.588±6e-3	0.716±6e-3	**0.0036±1e-4**	0.767±1e-2

Table 11: The comparison results of prediction on test sets (the test mode is 1, the observed views are 8, and query views are 1, 2). All test values are evaluated 5 times, recorded with mean and standard deviation.

Dataset	Query	Method	ARI-A↑	AMI-A↑	ARI-O↑	AMI-O↑	IoU↑	F1↑	MSE↓	OOA↑
CLEVR-SIMPLE	1	MulMON	0.681±1e-3	0.624±1e-3	**0.968±2e-3**	**0.962±2e-3**	**0.630±3e-3**	**0.751±4e-3**	**0.0014±3e-5**	N/A
		Ours	**0.819±3e-3**	**0.728±2e-3**	0.958±5e-3	0.960±4e-3	0.629±3e-3	0.754±3e-3	0.0017±8e-5	0.960±2e-2
	2	MulMON	0.714±1e-3	0.648±1e-3	**0.969±2e-3**	**0.960±1e-3**	**0.650±3e-3**	**0.767±3e-3**	**0.0013±2e-5**	N/A
		Ours	**0.834±4e-3**	**0.739±6e-3**	0.951±5e-3	0.951±5e-3	0.641±7e-3	0.763±6e-3	0.0017±5e-5	0.958±2e-2
CLEVR-COMPLEX	1	MulMON	0.591±7e-3	0.564±3e-3	0.948±2e-3	0.938±1e-3	**0.574±3e-3**	**0.703±3e-3**	**0.0020±1e-5**	N/A
		Ours	**0.769±1e-2**	**0.672±9e-3**	**0.951±5e-3**	**0.951±5e-3**	0.542±7e-3	0.678±8e-3	0.0025±2e-4	0.944±2e-2
	2	MulMON	0.613±9e-3	0.577±4e-3	0.945±2e-3	0.932±9e-4	**0.588±3e-3**	**0.716±3e-3**	**0.0020±2e-5**	N/A
		Ours	**0.788±1e-2**	**0.682±1e-2**	**0.950±6e-3**	**0.947±7e-3**	0.552±8e-3	0.689±8e-3	0.0024±1e-4	0.930±4e-2
SHOP-SIMPLE	1	MulMON	0.485±1e-2	0.568±5e-3	0.880±6e-3	0.873±3e-3	0.607±5e-3	0.729±5e-3	0.0050±1e-4	N/A
		Ours	**0.805±3e-3**	**0.735±3e-3**	**0.961±5e-3**	**0.959±4e-3**	**0.656±2e-3**	**0.769±2e-3**	**0.0035±6e-5**	0.866±3e-2
	2	MulMON	0.502±1e-2	0.581±5e-3	0.871±6e-3	0.862±3e-3	0.626±5e-3	0.745±5e-3	0.0050±1e-4	N/A
		Ours	**0.828±2e-3**	**0.751±2e-3**	**0.955±9e-4**	**0.951±2e-3**	**0.676±2e-3**	**0.786±3e-3**	**0.0035±2e-5**	0.851±1e-2
SHOP-COMPLEX	1	MulMON	0.707±9e-3	0.660±4e-3	0.878±5e-3	0.875±3e-3	**0.670±2e-3**	**0.776±2e-3**	0.0037±3e-5	N/A
		Ours	**0.779±8e-3**	**0.705±7e-3**	**0.944±7e-3**	**0.941±7e-3**	0.623±8e-3	0.739±8e-3	**0.0036±1e-4**	0.814±2e-2
	2	MulMON	0.730±9e-3	0.672±4e-3	0.860±6e-3	0.858±3e-3	**0.682±2e-4**	**0.787±6e-4**	**0.0038±3e-5**	N/A
		Ours	**0.791±7e-3**	**0.710±7e-3**	**0.938±8e-3**	**0.932±7e-3**	0.639±8e-3	0.756±9e-3	**0.0038±1e-4**	0.813±2e-2

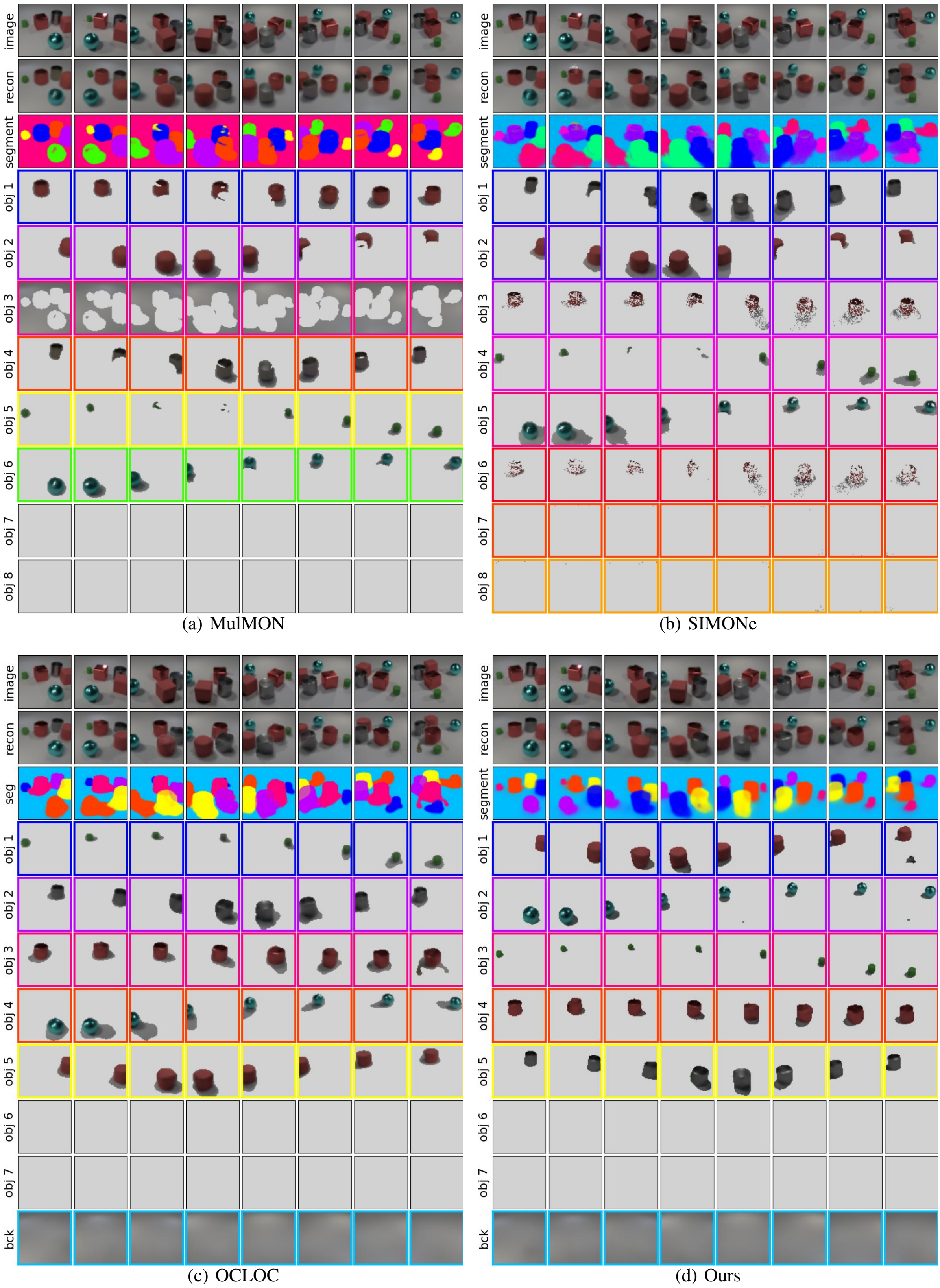
Table 12: The comparison results of prediction on test sets (the test mode is 2, the observed views are 8, and query views are 1, 2). All test values are evaluated 5 times, recorded with mean and standard deviation.

Dataset	Query	Method	ARI-A↑	AMI-A↑	ARI-O↑	AMI-O↑	IoU↑	F1↑	MSE↓	OOA↑
CLEVR-SIMPLE	1	MulMON	0.588±1e-3	0.549±1e-3	0.954±1e-3	0.949±1e-3	**0.570±2e-3**	**0.704±2e-3**	**0.0015±2e-5**	N/A
		Ours	**0.727±8e-3**	**0.638±7e-3**	**0.960±6e-3**	**0.960±5e-3**	0.546±7e-3	0.687±9e-3	0.0019±2e-4	0.953±2e-2
CLEVR-COMPLEX	1	MulMON	0.480±1e-2	0.477±5e-3	0.910±2e-3	0.896±2e-3	**0.512±3e-3**	**0.654±3e-3**	**0.0023±4e-5**	N/A
		Ours	**0.681±7e-3**	**0.593±5e-3**	**0.951±4e-3**	**0.952±3e-3**	0.491±3e-3	0.639±4e-3	0.0025±8e-5	0.890±2e-2
SHOP-SIMPLE	1	MulMON	0.413±6e-3	0.516±2e-3	0.872±5e-3	0.862±2e-3	0.567±2e-3	0.702±2e-3	0.0058±1e-4	N/A
		Ours	**0.752±6e-3**	**0.684±4e-3**	**0.956±2e-3**	**0.956±2e-3**	**0.608±3e-3**	**0.733±3e-3**	**0.0046±4e-4**	**0.823±2e-2**
SHOP-COMPLEX	1	MulMON	0.648±5e-3	0.609±3e-3	0.869±8e-3	0.866±4e-3	**0.628±4e-3**	**0.746±3e-3**	0.0042±5e-5	N/A
		Ours	**0.728±1e-2**	**0.657±9e-3**	**0.930±7e-3**	**0.935±5e-3**	0.574±8e-3	0.701±8e-3	**0.0040±1e-4**	0.811±2e-2

Table 13: The comparison results of prediction on test sets (the test mode is 1, the observed views are 9, and query views are 1). All test values are evaluated 5 times, recorded with mean and standard deviation.

Dataset	Query	Method	ARI-A↑	AMI-A↑	ARI-O↑	AMI-O↑	IoU↑	F1↑	MSE↓	OOA↑
CLEVR-SIMPLE	1	MulMON	0.748±5e-4	0.676±7e-4	**0.969±2e-3**	**0.962±1e-3**	**0.675±1e-3**	**0.785±2e-3**	**0.0011±6e-6**	N/A
		Ours	**0.858±1e-2**	**0.779±1e-2**	0.959±8e-3	0.960±7e-3	0.668±1e-2	0.778±1e-2	0.0016±2e-4	0.967±2e-2
CLEVR-COMPLEX	1	MulMON	0.640±8e-3	0.597±4e-3	0.944±2e-3	0.936±2e-3	**0.603±3e-3**	**0.728±3e-3**	**0.0020±6e-6**	N/A
		Ours	**0.817±4e-3**	**0.716±4e-3**	**0.963±4e-3**	**0.962±3e-3**	0.571±3e-3	0.705±2e-3	0.0022±5e-5	0.955±2e-2
SHOP-SIMPLE	1	MulMON	0.551±2e-2	0.615±9e-3	0.870±4e-3	0.871±2e-3	0.648±5e-3	0.763±5e-3	0.0049±1e-4	N/A
		Ours	**0.848±7e-3**	**0.778±7e-3**	**0.952±5e-3**	**0.951±4e-3**	**0.691±6e-3**	**0.795±5e-3**	**0.0038±3e-4**	0.824±3e-2
SHOP-COMPLEX	1	MulMON	0.755±4e-3	0.694±1e-3	0.847±7e-3	0.853±4e-3	**0.700±3e-3**	**0.800±4e-3**	**0.0039±6e-5**	N/A
		Ours	**0.804±2e-2**	**0.733±1e-2**	**0.938±9e-3**	**0.938±7e-3**	0.655±1e-2	0.765±1e-2	0.0041±4e-4	0.812±7e-2

Table 14: The comparison results of prediction on test sets (the test mode is 2, the observed views are 9, and query views are 1). All test values are evaluated 5 times, recorded with mean and standard deviation.



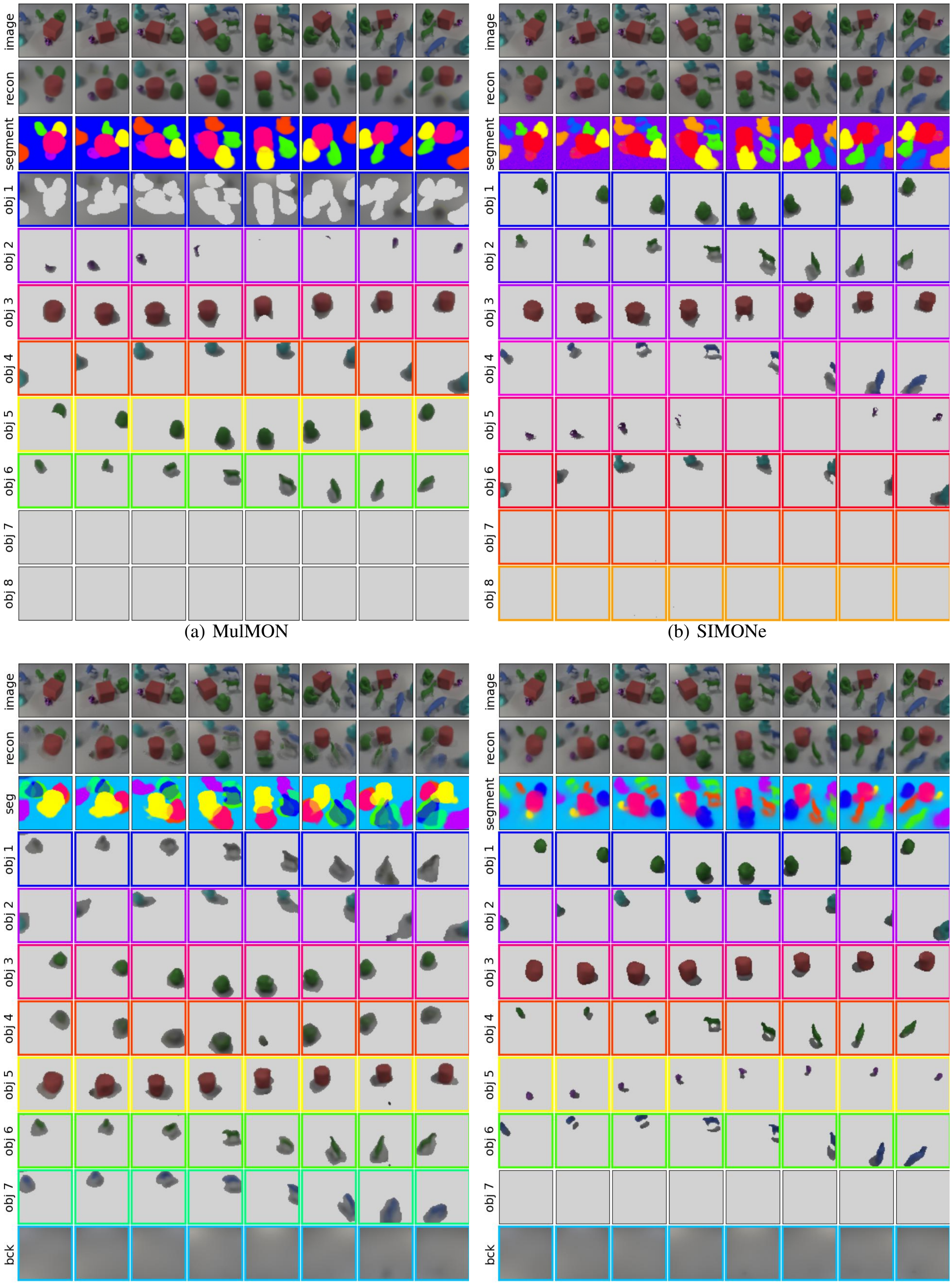
(a) MulMON

(b) SIMONe

(c) OCLOC

(d) Ours

Figure 3: Qualitative comparison of observation on the CLEVR-SIMPLE dataset, observed views are 8



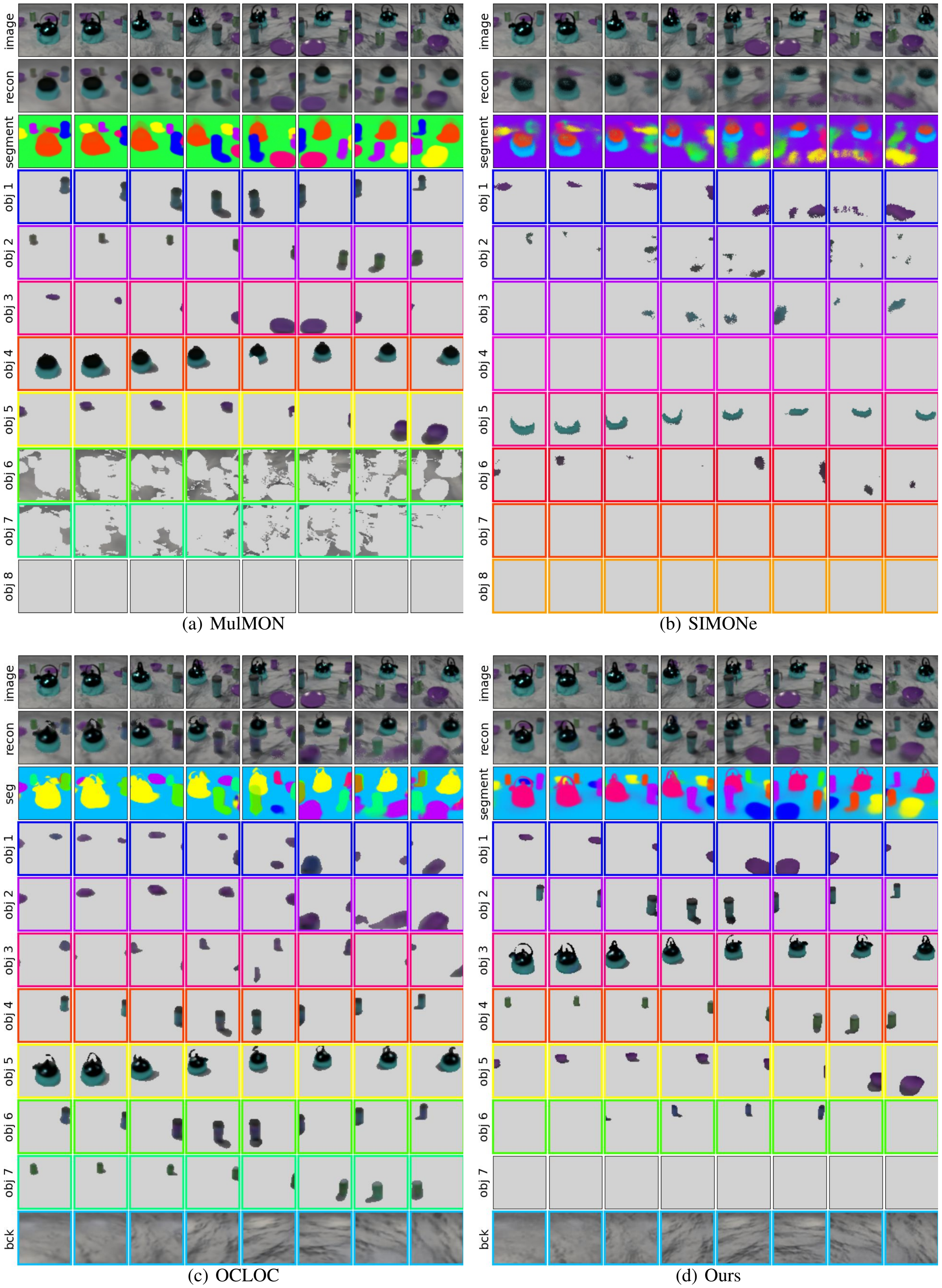
(a) MulMON

(b) SIMONe

(c) OCLOC

(d) Ours

Figure 4: Qualitative comparison of observation on the CLEVR-COMPLEX dataset, observed views are 8



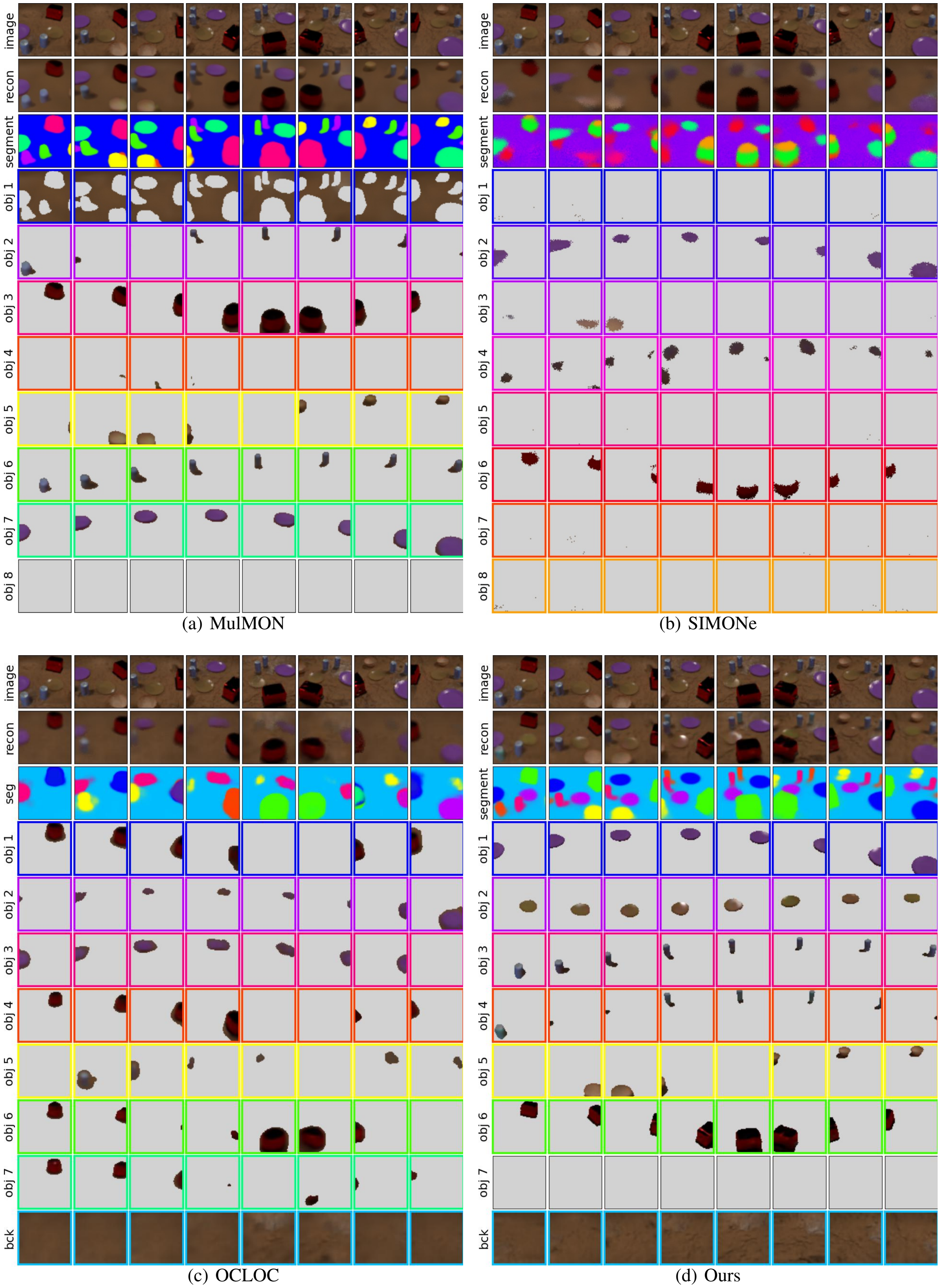
(a) MulMON

(b) SIMONe

(c) OCLOC

(d) Ours

Figure 5: Qualitative comparison of observation on the SHOP-SIMPLE dataset, observed views are 8



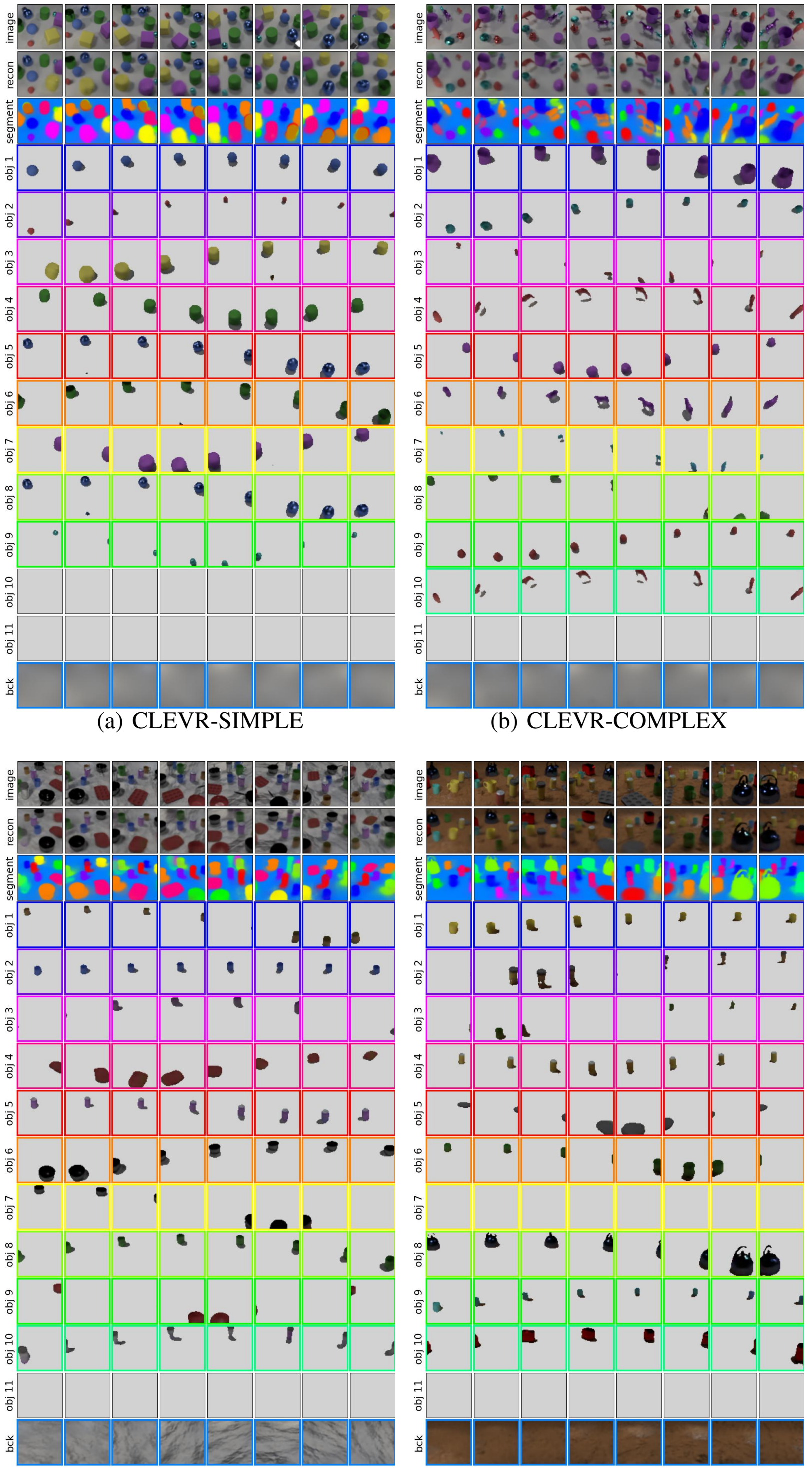
(a) MulMON

(b) SIMONe

(c) OCLOC

(d) Ours

Figure 6: Qualitative comparison of observation on the SHOP-COMPLEX dataset, observed views are 8



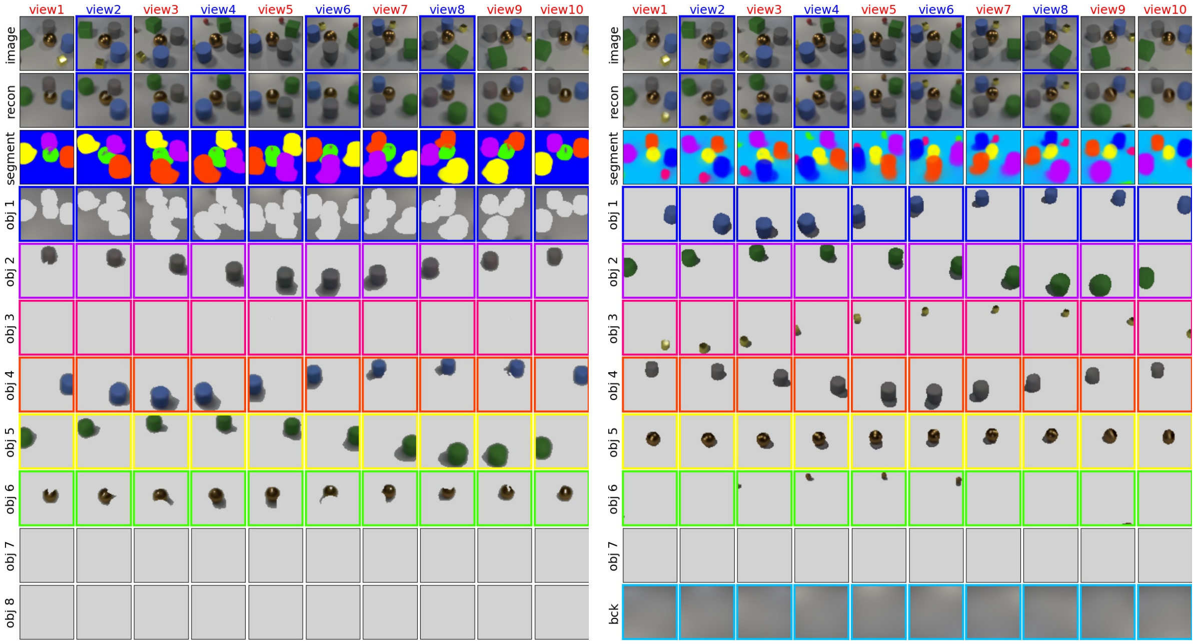
(a) CLEVR-SIMPLE

(b) CLEVR-COMPLEX

(c) SHOP-SIMPLE

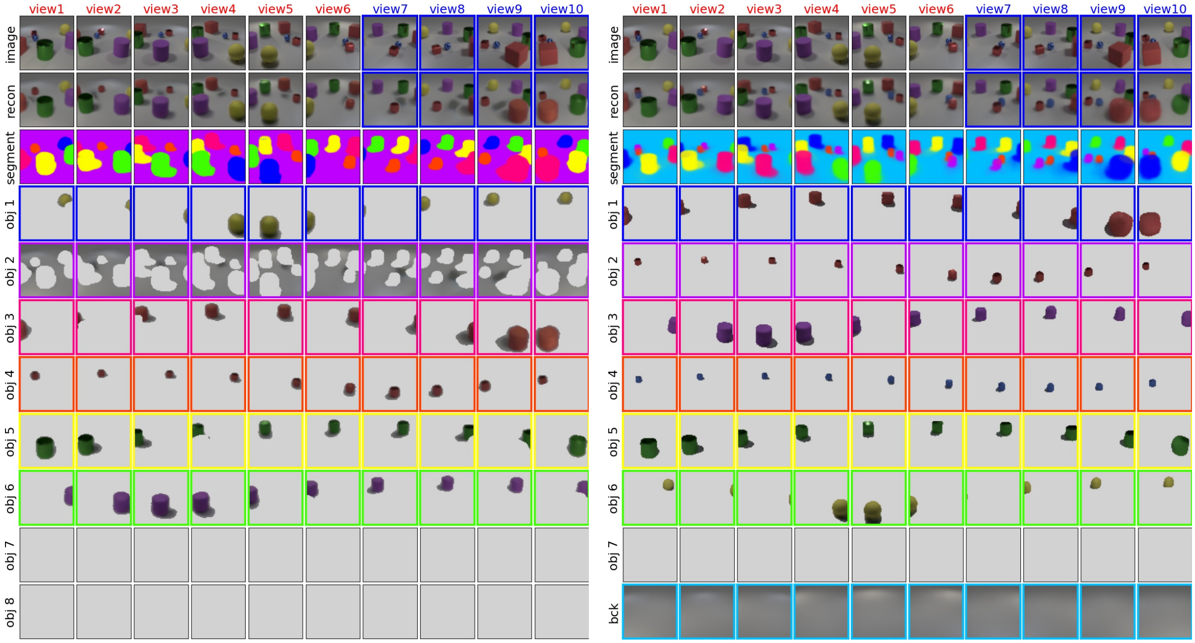
(d) SHOP-COMPLEX

Figure 7: Visualization results on the general sets of dataset, observed views are 8



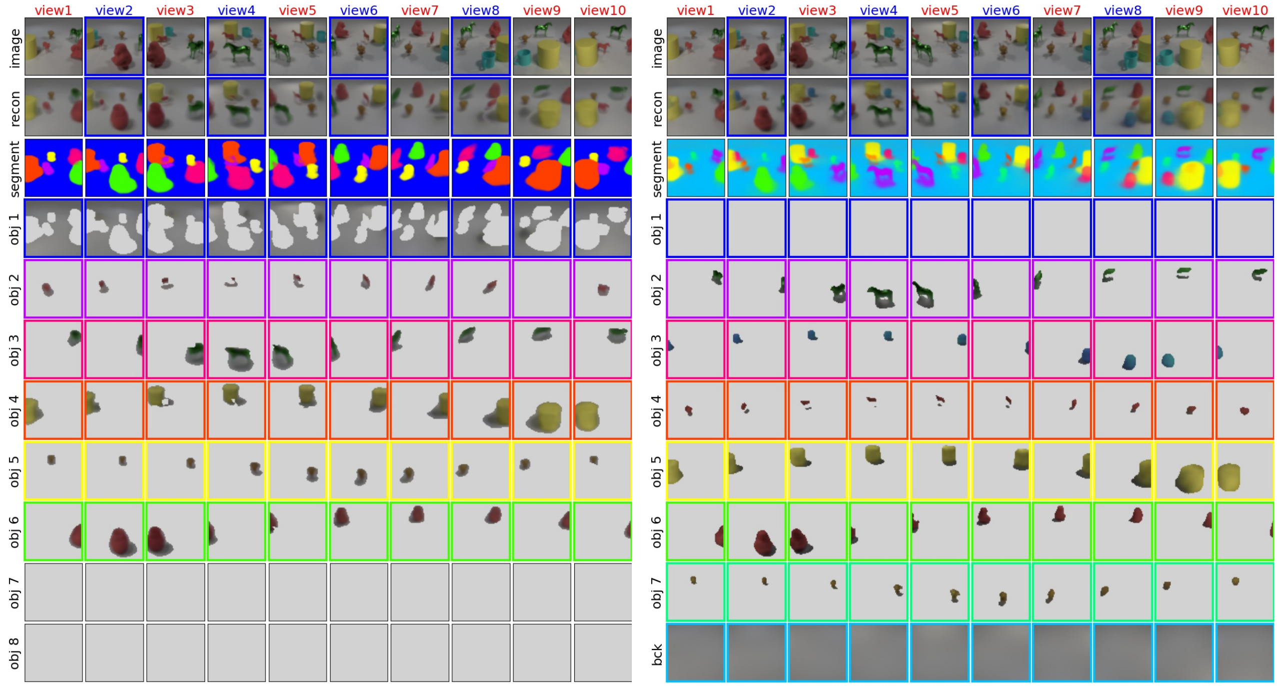
(a) MulMON (b) Ours

Figure 8: Qualitative comparison of prediction on the CLEVR-SIMPLE dataset. The observed views are 6, test mode is 1, query views are 4.



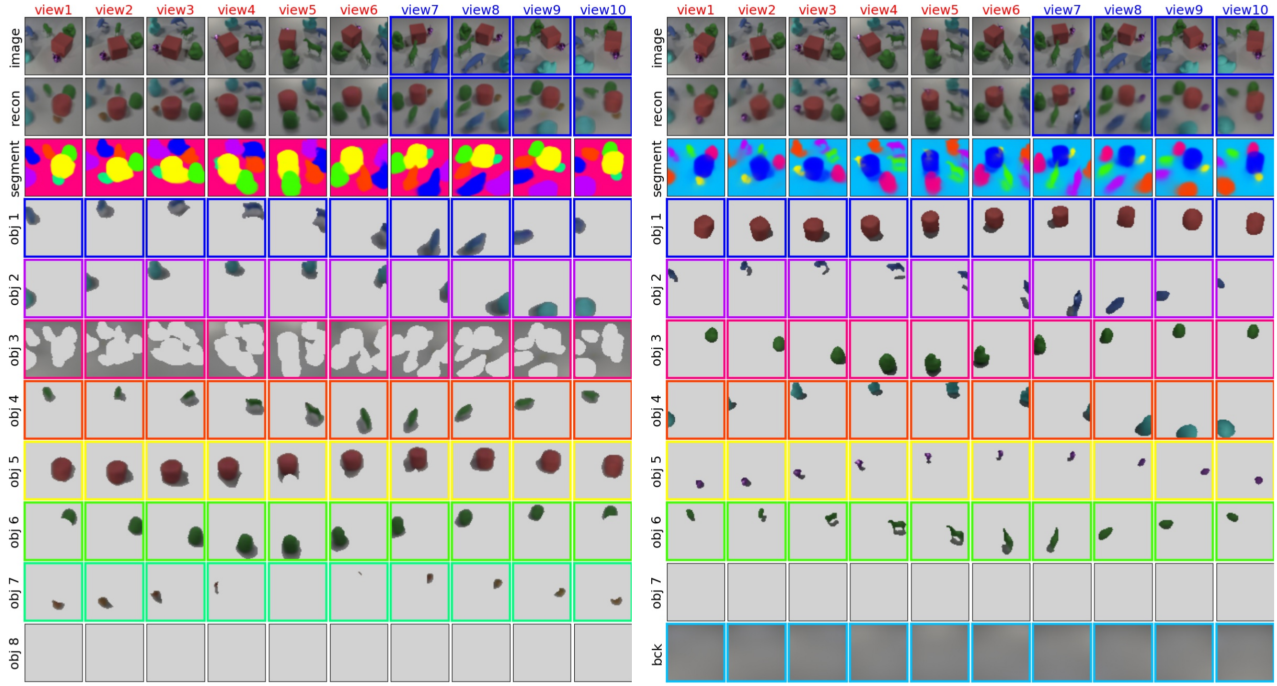
(a) MulMON (b) Ours

Figure 9: Qualitative comparison of prediction on the CLEVR-SIMPLE dataset. The observed views are 6, test mode is 2, query views are 4.



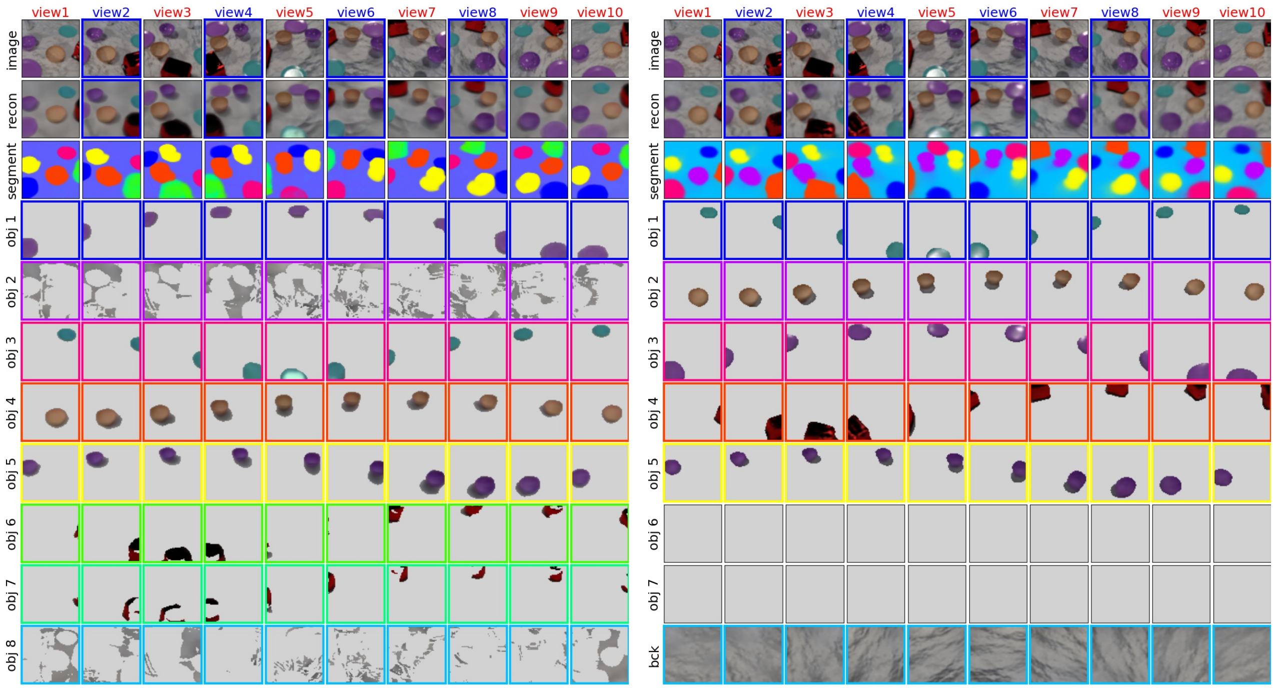
(a) MulMON (b) Ours

Figure 10: Qualitative comparison of prediction on the CLEVR-COMPLEX dataset. The observed views are 6, test mode is 1, query views are 4.



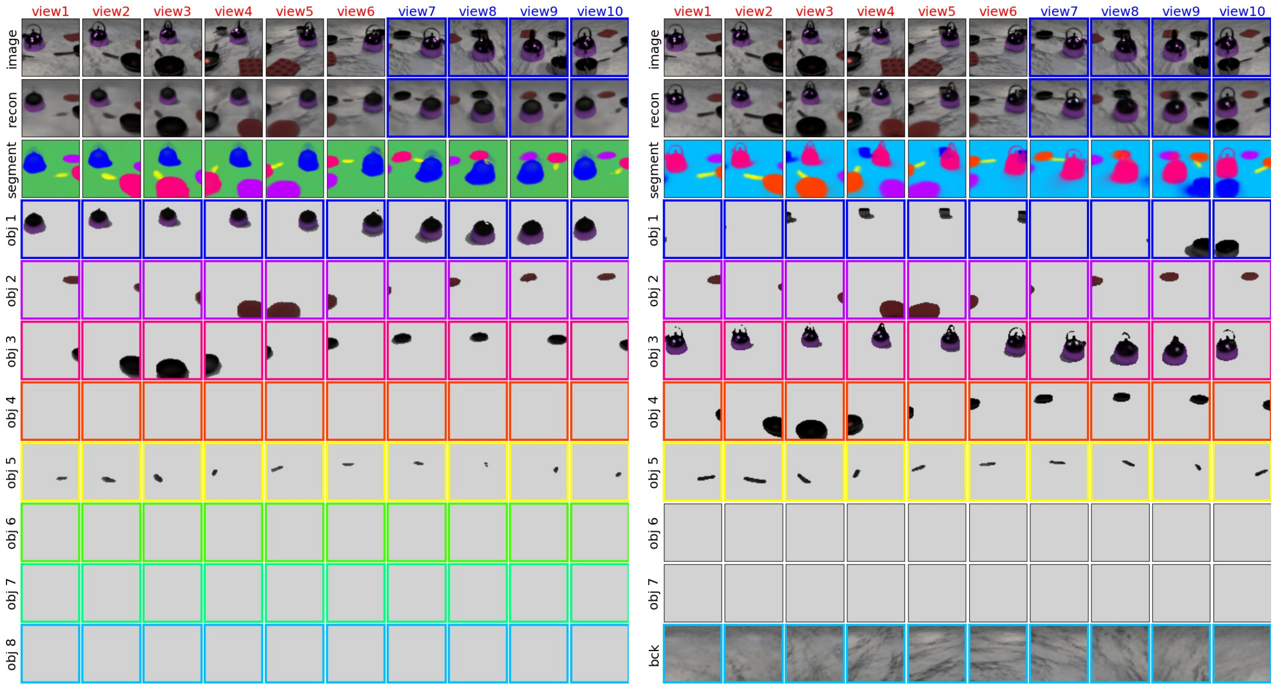
(a) MulMON (b) Ours

Figure 11: Qualitative comparison of prediction on the CLEVR-COMPLEX dataset. The observed views are 6, test mode is 2, query views are 4.



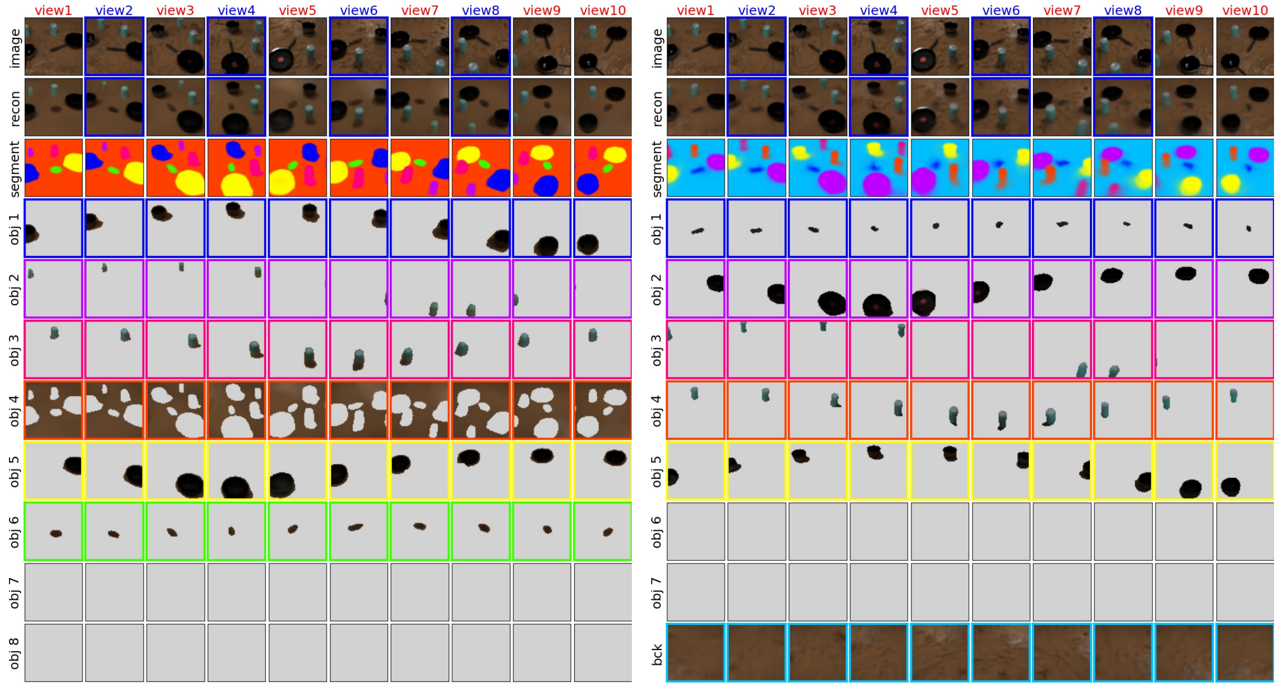
(a) MulMON (b) Ours

Figure 12: Qualitative comparison of prediction on the SHOP-SIMPLE dataset. The observed views are 6, test mode is 1, query views are 4.



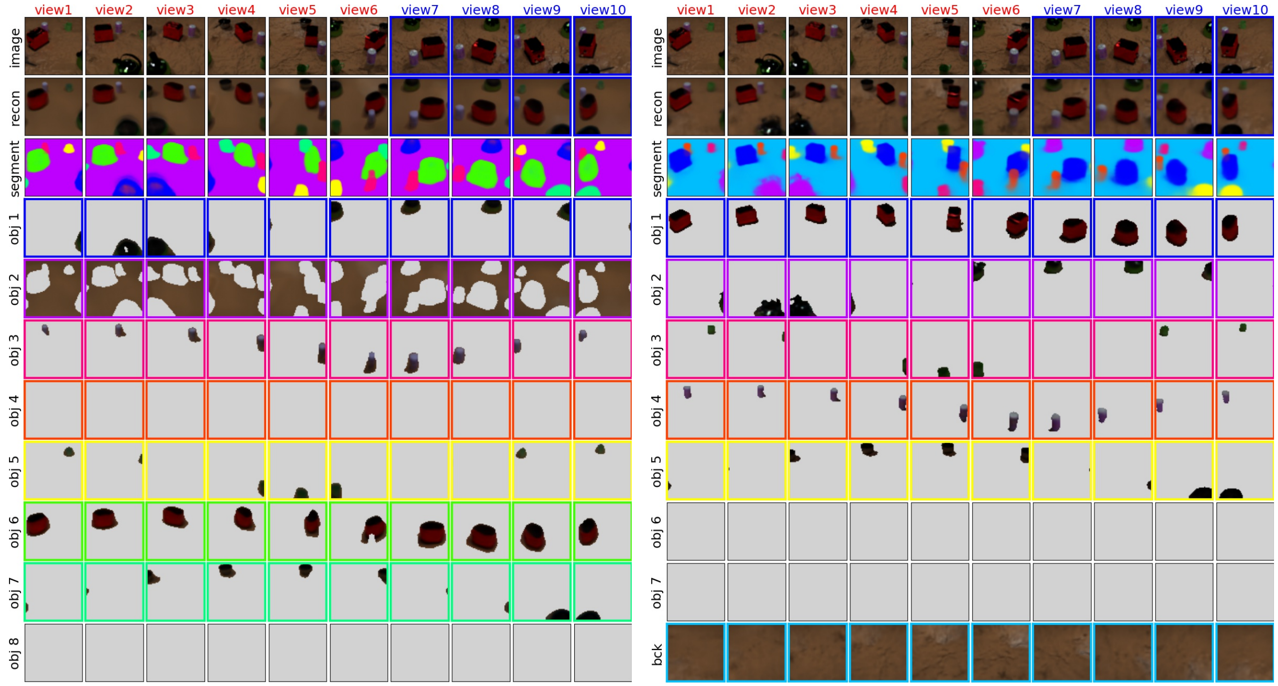
(a) MulMON (b) Ours

Figure 13: Qualitative comparison of prediction on the SHOP-SIMPLE dataset. The observed views are 6, test mode is 2, query views are 4.



(a) MulMON (b) Ours

Figure 14: Qualitative comparison of prediction on the SHOP-COMPLEX dataset. The observed views are 6, test mode is 1, query views are 4.



(a) MulMON (b) Ours

Figure 15: Qualitative comparison of prediction on the SHOP-COMPLEX dataset. The observed views are 6, test mode is 2, query views are 4.

References

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006.

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.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

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.

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.

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.

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.

Nguyen Xuan, Vinh Julien, South Wales, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 2010.

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, 2019.

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, 2022.