Efficient Object-Centric Representation Learning using Masked Generative Modeling

Anonymous authors
Paper under double-blind review

Abstract

Learning object-centric representations from visual inputs in an unsupervised manner have drawn focus to solve more complex tasks, such as reasoning and reinforcement learning. However, current state-of-the-art methods, relying on autoregressive transformers or diffusion models to generate scenes from object-centric representations, suffer from computational inefficiency due to their sequential or iterative nature. This computational bottleneck limits their practical application and hinders scaling to more complex downstream tasks. To overcome this, we propose MOGENT, an efficient object-centric learning framework based on masked generative modeling. MOGENT conditions a masked bidirectional transformer on learned object slots and employs a parallel iterative decoding scheme to generate scenes, enabling efficient compositional generation. We conduct experiments on 3D Shapes and CLEVR, demonstrating that MOGENT significantly improves computational efficiency, accelerating the generation process by up to 10x compared to autoregressive models. Importantly, the efficiency is attained followed by a strong or competitive performance on object segmentation and compositional generation tasks.

1 Introduction

A key aspect of human intelligence the ability to perceive their surroundings as composition of objects and their relationships (Spelke, 1990; 2013). Such abstraction allows humans to flexibly generalize and reason about novel scenarios by composing existing conceptual knowledge, a capability known as compositional generalization in machine learning (Greff et al., 2020; Goyal & Bengio, 2022; Lake et al., 2017). Inspired by this ability, the field of object-centric learning aims to develop models that can decompose complex scenes, such as images or videos, into individual object representations in an unsupervised manner. A prevalent approach in object-centric learning is to represent each object in an image or video as a set of representations, often referred to as “slots” (Burgess et al., 2019; Greff et al., 2019; Locatello et al., 2020). Early works achieved object discovery and disentanglement by introducing various inductive biases, such as grouping nearby pixels (Greff et al., 2017; van Steenkiste et al., 2018), modeling object properties (e.g. position, size, depth, etc.) explicitly (Eslami et al., 2016; Jiang et al., 2020), and modeling foreground and background separately (Lin et al., 2020b;a).

Among various architectures, Slot Attention (Locatello et al., 2020) emerged as an influential method, employing iterative attention (Vaswani et al., 2017) over encoded features to bind information into distinct object slots and reconstructing the scene using a mixture-based decoder (Watters et al., 2019). Trained on a simple input reconstruction objective, Slot Attention is a popular architectural choice and has been extended for both images (Singh et al., 2022a; Seitzer et al., 2023; Didolkar et al., 2025) and videos (Singh et al., 2022b; Kipf et al., 2022; Zadaianchuk et al., 2023; Wu et al., 2023b). Recent works have focused on enhancing the generative capabilities of these slot-based approaches. These models often leverage the extracted slots as conditional inputs for sequential generators such as autoregressive transformers or diffusion models such as Latent Diffusion Model (LDM) (Rombach et al., 2022), enabling object-centric disentanglement and compositional generation on more realistic datasets (Singh et al., 2022a;b; Wu et al., 2023b; Jiang et al., 2023; Kakogeorgiou et al., 2024).

However, these models often suffer from computational inefficiency. For example, when using an autoregressive transformer, generation requires (# of patches) steps per image (Figure 1 (a)). This is especially challenging in object-centric learning, as effective object-centric disentanglement typically requires smaller patches to capture objects of varying sizes and partial occlusions accurately. On the other hand, diffusion-based models significantly reduce the number of generation steps required per image, yet remain computationally expensive memory-wise and time-wise due to their iterative refinement procedure (Wu et al., 2023b; Jiang et al., 2023). While their high computational cost have been pointed out as a limitation by Wu et al. (2023b), no prior works have worked on improving efficiency of object-centric generation.

In this work, we present MOGENT (**M**asked **O**bject-centric **GEN**erative **T**ransformer), an object-centric learning framework that leverages the efficiency of masked generative modeling. MOGENT employs a masked bidirectional transformer decoder, conditioned on extracted slots, to predict masked visual tokens representing image patches. Inspired by the success of parallel decoding schemes on generating images and videos (Chang et al., 2022a; Yu et al., 2023; 2024), we utilize the iterative refinement scheme based on MaskGIT (Chang et al., 2022a). This allows MOGENT to generate a large number of tokens in parallel at each step, significantly improving computational efficiency. For example, generating a 128x128 image requires only 20 decoding steps with MOGENT, a substantial reduction from the 1024 steps typically required by autoregressive baselines (Singh et al., 2022a). More importantly, MOGENT can generate efficiently regardless of the image resolution as the number of steps to generate does not depend on the image resolution. We further show that integrating the Query Slot Attention (QSA) (Jia et al., 2023) to extract slots and adjusting the initialization and loss function of the transformer is crucial for achieving effective object-centric representation learning within this efficient framework. Experiments on the 3D Shapes (Burgess et al., 2019) and CLEVR (Johnson et al., 2017) datasets show that MOGENT achieves up to a 10x speedup in both training and inference compared to relevant baselines. Notably, this efficiency gain is realized without compromising, and often improving upon, representation learning and generation quality.

2 Related Works

Object-centric Learning. Learning to represent objects in the scene using "slots" (Locatello et al., 2020; Burgess et al., 2019; Greff et al., 2019) has been long explored in the literature. A key inductive bias to achieve object-centric disentanglement is iterative inference, which has been achieved by applying iteration over objects (Eslami et al., 2016; Burgess et al., 2019) or iterative refinement (Greff et al., 2019; Locatello et al., 2020; Goyal et al., 2021). Additionally, adding further priors about object properties (e.g. position, size, depth, etc.) (Eslami et al., 2016; Jiang et al., 2020), foreground and background (Lin et al., 2020a;b), or temporal indifference (Hsieh et al., 2018; Nakano et al., 2023), has also been found to be effective in improving object-centric disentanglement. Slot Attention (Locatello et al., 2020) is one of the commonly-used model, which uses iterative attention mechanism (Vaswani et al., 2017) and mixture-based decoder (Watters et al., 2019) to learn slot representations from various datasets. However, the efficacy of these early object-centric models is often limited when dealing with complex real-world scenes (Yang & Yang, 2022). To address this, several works have explored improving Slot Attention, such as employing bi-level optimization (Chang et al., 2022b; Jia et al., 2023), adding spatial locality prior (Chakravarthy et al., 2023), or learning quantized slot representations (Singh et al., 2023; Wu et al., 2024).

Other works have explored improving generation performance of slot-based models by replacing the mixture-based decoder with models with higher capacity, such as transformer or diffusion models (Singh et al., 2022a;b; Sajjadi et al., 2022; Wu et al., 2023a;b; Jiang et al., 2023; Kakogeorgiou et al., 2024). For example, SLATE (Singh et al., 2022a) and STEVE (Singh et al., 2022b) use an autoregressive transformer to generate images or videos from slots, respectively. They train a discretized VAE (Im et al., 2017) to tokenize the inputs, extract slots using Slot Attention, and generate scenes using a slot-conditioned transformer decoder. In contrast, SlotDiffusion (Wu et al., 2023b) employs LDM (Rombach et al., 2022) to generate scenes using slot-conditioned denoising within the latent space of a pretrained VQ-VAE (Van Den Oord et al., 2017). While powerful, both the sequential token prediction in autoregressive models and the iterative nature of diffusion lead to computational inefficiency and slow generation times (Wu et al., 2023b; Jiang et al., 2023). Addressing this bottleneck, our work employs masked generative modeling, aiming for more efficient object-centric learning and parallelizable generation. Given this focus on improving generation efficiency

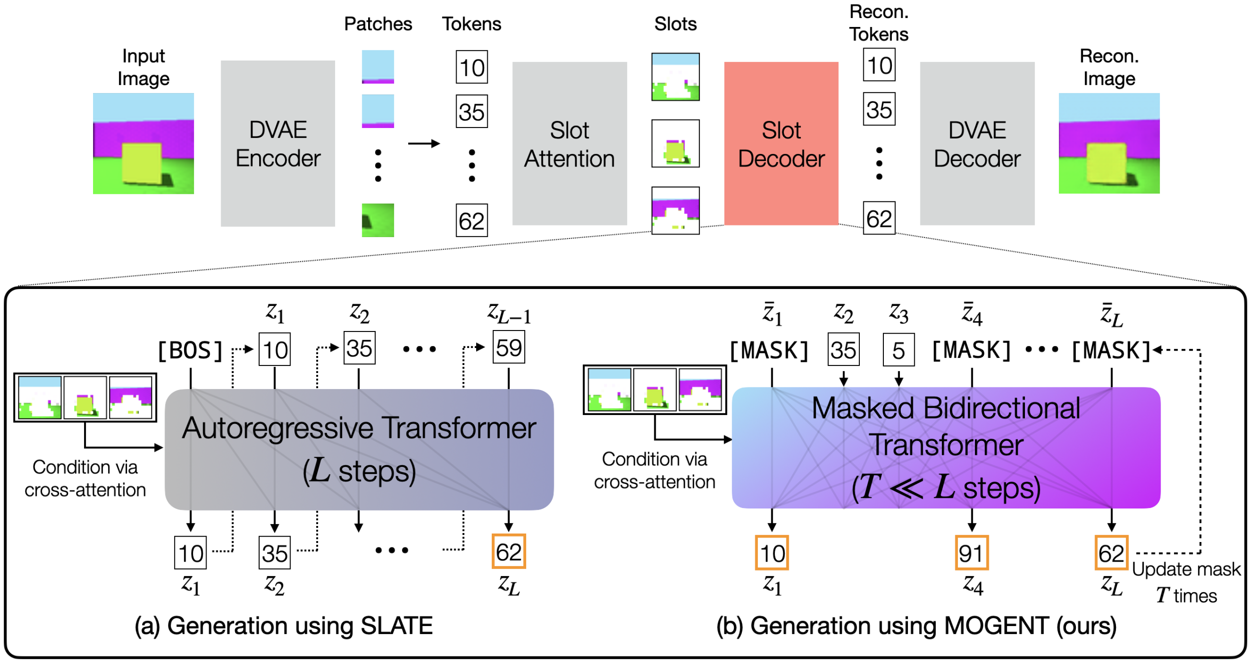


Figure 1: Overview of the generation process using (a) SLATE and (b) MOGENT. As SLATE uses an autoregressive transformer for slot-to-token decoding, generation takes as many steps as the number of tokens (L) to represent a single image. On the other hand, by employing a masked bidirectional transformer, MOGENT can generate tokens in parallel, reducing the number of steps to generate by a large margin.

via a non-sequential strategy, we primarily compare our approach against SLATE, as its autoregressive slot-to-token decoder serves as a key contrasting baseline for generation methodology.

Masked Generative Modeling. Autoregressive decoding is known to suffer from the slow inference speed and sequential error accumulation, and have been extensively studied in the field of natural language processing. Non-autoregressive generation algorithms has emerged to address the challenges of autoregressive decoding, with masked token prediction recognized as a variant of this approach (Devlin et al., 2019; Ghazvininejad et al., 2019; Mansimov et al., 2019). Application to images has also been explored (Chang et al., 2022a; Lee et al., 2022), in which MaskGIT (Chang et al., 2022a) improved both image generation quality and efficiency. MaskGIT has been extended to video prediction (Yan et al., 2023), text-to-image (Chang et al., 2023; Patil et al., 2024), text-to-video (Yu et al., 2023; 2024; Villegas et al., 2023), multi-modal generation (Chang et al., 2023; Kim et al., 2023; Mizrahi et al., 2024), LiDAR point generation (Zhang et al., 2018), motion generation (Guo et al., 2024; Pinyoanuntapong et al., 2024b;a), and neural simulation of interactive environments (Bruce et al., 2024). Our work is the first to investigate and adapt masked generative modeling to object-centric representation learning.

3 Method

We begin by reviewing the background of object-centric learning using SLATE (Singh et al., 2022a) (Section 3.1), which we build our model on. We then detail our object-centric masked generative transformer architecture and explain how to perform compositional image generation using the learned representations of MOGENT. (Section 3.2). Figure 1 illustrates the architecture of SLATE and MOGENT.

3.1 Preliminary: Slot-based object-centric learning using SLATE

The goal of object-centric learning is to learn a set of representations, or “slots”, that each correspond to an object within a scene. A commonly-used architecture is Slot Attention (Locatello et al., 2020), which learns slot representations by computing iterative attention between randomly initialized slots and encoded input image. SLATE (Singh et al., 2022a) extends this work by combining Discrete VAE (DVAE) (Im et al., 2017) and an autoregressive transformer decoder (Vaswani et al., 2017).

Specifically, SLATE encodes an input image $\mathbf{x}$ through the DVAE encoder f_ϕ, to produce log probabilities, $\mathbf{o}$, for a categorical distribution with V classes. A "soft" one-hot encoding $\mathbf{z}_{\text{soft}}$ is sampled from a relaxed categorical distribution (Jang et al., 2017), and decoded via the DVAE decoder, g_θ. Denoting the temperature of the relaxed categorical distribution as τ_{DVAE}, the entire process can be written as

$$\tilde{\mathbf{x}} = g_\theta(\mathbf{z}_{\text{soft}}) \ \text{ where } \ \mathbf{z}_{\text{soft}} \sim \text{RelaxedCategorical}(\mathbf{o}; \tau_{\text{DVAE}}), \ \ \mathbf{o} = f_\phi(\mathbf{x}). \tag{1}$$

To compute slots, the tokens from the DVAE encoder are first mapped to embeddings, $\mathbf{e}$, using a learned dictionary. Learned positional embeddings, $\mathbf{p}_\phi$, are added to the embeddings to incorporate positional information of the tokens. Then, the embeddings are fed to Slot Attention (Locatello et al., 2020) encoder to extract K slots, $\mathbf{s}_{1:K}$. This process can be written as,

$$\mathbf{s}_{1:K} = \text{SlotAttention}(\mathbf{e}) \ \text{ where } \ \mathbf{e} = \text{Dictionary}_\phi(\mathbf{z}) + \mathbf{p}_\phi, \ \ \mathbf{z} \sim \text{Categorical}(\mathbf{o}). \tag{2}$$

Finally, starting from a `[BOS]` token, an autoregressive transformer (Vaswani et al., 2017), p_θ, decodes the slots back into the discrete tokens one at a time, which can be formulated as generation using next token prediction:

$$p_\theta(z_1, \cdots, z_L | \mathbf{s}_{1:K}) = \prod_{l=1}^{L} p_\theta(z_l | z_1, \cdots, z_{l-1}, \mathbf{s}_{1:K}), \tag{3}$$

where L denotes the number of tokens. The resulting tokens can be decoded back into an image by the DVAE decoder, g_θ, enabling compositional scene generation.

Overall, DVAE is trained to minimize the negative log-likelihood, $\mathcal{L}_{\text{DVAE}} = \mathbb{E}_{\mathbf{z}_{\text{soft}}}[-\log g_\theta(\mathbf{x}|\mathbf{z}_{\text{soft}})]$, using reconstruction loss. Slot Attention and the transformer decoder are trained to minimize the negative log-likelihood, $\mathcal{L}_{\text{ST}} = \mathbb{E}_{\mathbf{s}_{1:K}}[-\sum_{l=1}^{L} \log p_\theta(z_l|z_1, \cdots, z_{l-1}, \mathbf{s}_{1:K})]$, using cross-entropy loss. The entire model is trained together. Please refer to Singh et al. (2022a) for more information on training details.

3.2 MOGENT

As explained in the previous section, most object-centric generative models generate new scenes by first inferring the slot representations and then decoding them back to the pixel space. While the choice of the slot-to-token decoder is important for both effective object-centric disentanglement and high generation quality (Wu et al., 2023a), existing options present notable trade-offs. Mixture-based decoders (Watters et al., 2019) are computationally efficient but often possess limited capacity, tending to produce blurry results on complex data (Singh et al., 2022a; Wu et al., 2023a). In contrast, decoders with higher capacity such as autoregressive transformer-based (Singh et al., 2022a;b; Kakogeorgiou et al., 2024) or latent diffusion model-based decoders (Wu et al., 2023b; Jiang et al., 2023) generate higher-quality images but are computationally expensive, hindering training and inference speed..

To address these limitations, we propose MOGENT, a framework leveraging masked generative modeling for object-centric representation learning and efficient compositional generation. Drawing inspiration from the success of masked generative modeling in generating high-quality images and videos with high efficiency (Chang et al., 2022a; Yu et al., 2023; 2024), we utilize the MaskGIT (Chang et al., 2022a) framework as the decoder. Specifically, we view the slot-to-token decoding problem as generation using masked token prediction by replacing the autoregressive transformer of SLATE with BERT (Devlin et al., 2019), a transformer with bidirectional attention.

During training, the bidirectional transformer is trained to predict the masked parts of the input tokens. A binary mask, $\mathbf{m}(r) = [m_l]_{l=1}^{L}$, is generated using a predefined masking scheduler function, $\gamma(r) \in (0, 1]$ as follows: first sample a ratio, r, from a uniform distribution, $\mathcal{U}(0, 1)$, then uniformly select $\lceil \gamma(r) \cdot L \rceil$ tokens to mask out of L total tokens. Following MaskGIT, we choose cosine function as the masking scheduler.

The token, z_l, is replaced with a `[MASK]` token if $m_l = 1$, otherwise unmasked. Denoting the masked input $\bar{\mathbf{z}}^r = \mathbf{z} \odot \mathbf{m}(r)$, we train the bidirectional transformer, p_θ to minimize the negative log-likelihood of the masked tokens using cross-entropy loss:

$$\mathcal{L}_{\mathrm{ST}} = \mathbb{E}_{\mathbf{s}_{1:K}} \left[\mathbb{E}_{\mathbf{m}(r) \sim p_{\mathcal{U}}} \left[- \sum_{l=1}^{L} \log p_\theta(z_l | \bar{\mathbf{z}}^r, \mathbf{s}_{1:K}) \right] \right]. \tag{4}$$

Similar to SLATE, we incorporate cross-attention layers in the bidirectional transformer for slot conditioning.

For inference, we use the iterative parallel decoding scheme of MaskGIT. We start with a blank canvas with all tokens masked out and operate the following procedures iteratively for T steps; (1) Predict the probabilities for all the masked tokens at step t, $\bar{\mathbf{z}}^{<t} = \mathbf{z} \odot \mathbf{m}^t$. (2) Sample a token based on the predicted probabilities. (3) Compute the number of tokens to mask using the mask scheduler function. (4) Decide tokens to unmask for the next iteration, $\bar{\mathbf{z}}^t$ using the schedule from (3) and the log probabilities from (1) used as "confidence" score. As $\gamma(r) = \gamma(t/T)$ is a monotonically decreasing function, the iterative decoding scheme ensures that the number of unmasked tokens monotonically increase until all tokens are generated at step T.

Formally, the iterative decoding scheme can be viewed as a generation using "next set-of-tokens prediction" (Li et al., 2024). Let $\mathcal{S}$ be an ordered list expressing the schedule of unmasking by the scheduler function, $\mathcal{S} = [\bar{\mathbf{z}}^1, \bar{\mathbf{z}}^2, \cdots, \bar{\mathbf{z}}^T]$. Note that $\{\bar{\mathbf{z}}^t\}_{t \in \{1, \cdots, T\}}$ do not contain any overlapping tokens and is complete. Then, the generation using the bidirectional transformer can be expressed as,

$$p_\theta(z_1, \cdots, z_L | \mathbf{s}_{1:K}) = \prod_{t=1}^{T} p(\bar{\mathbf{z}}^t | \bar{\mathbf{z}}^{<t}) = p(\bar{\mathbf{z}}^1) p(\bar{\mathbf{z}}^2 | \bar{\mathbf{z}}^1) p(\bar{\mathbf{z}}^3 | \bar{\mathbf{z}}^1, \bar{\mathbf{z}}^2) \cdots p(\bar{\mathbf{z}}^T | \bar{\mathbf{z}}^{<T}). \tag{5}$$

Therefore, the generation requires T steps in total, which is non-dependent of the number of tokens, L.

Empirically, we find that naively replacing the original slot decoder of SLATE with a masked bidirectional transformer is insufficient for learning object-centric representations. We hypothesize this stems from differences in how spatial locality priors—the assumption that nearby pixels often belong to the same object (Chakravarthy et al., 2023)—are handled. While SLATE's sequential autoregressive decoding naturally focuses on local neighborhoods, the bidirectional attention in MOGENT allows attending to further away tokens during decoding, promoting a more global attention but potentially weakening this implicit spatial locality bias.

To mitigate this, we make the following three changes to the model architecture and training setup. First, we adopt Query Slot Attention (QSA) which uses learnable query initializations instead of random initialization. As shown by Chang et al. (2022b); Jia et al. (2023), using random initialization plays a minimal role and could be removed. We empirically find that using QSA leads to large improvements in object-centric disentanglement for our framework. Secondly, we initialize the mask embeddings as 0 (*zero mask init*). This simple technique facilitates the model's ability to differentiate between the slots conditioning and the masked tokens, especially during early training, contributing to more stable learning and better disentanglement. Thirdly, while most previous works on masked generative modeling implement the loss function as the cross-entropy loss on both masked and unmasked tokens, we train MOGENT using cross-entropy loss on only the masked tokens, as derived in Equation 4. This objective discourages the model to learn an identity map on the unmasked tokens, and encourages it to use information from surrounding unmasked tokens to predict the masked tokens. We reason that this loss formulation not only prevents codebook collapse, but also motivates the model to capture the semantic information in the image for better disentanglement. We summarize our empirical findings regarding the model architecture in Section 4.4.

Compositional Generation. The learned slots each represent the individual objects in the image. Therefore, following Singh et al. (2022a); Wu et al. (2023b), we can build a library of the representations from the extracted slots. Then, we can generate images with novel combinations of objects by composing the representations ("concepts") from the library.

As described by Singh et al. (2022a); Wu et al. (2023b), we can generate new images compositionally via the following steps: (1) Collect slots from all training images. (2) Apply K-means clustering to find K concepts

Table 1: Comparison of computation requirements of SLATE and MOGENT using 3D Shapes dataset. All metrics were computed on a single NVIDIA Tesla V100 GPU, with batch size of 64 for training and 1 for test.

		SLATE	MOGENT (Ours)
Train	# of parameters	3.6M	3.7M
	Time [s]	0.465	**0.056**
Test	Time [s]	1.929	**0.182**

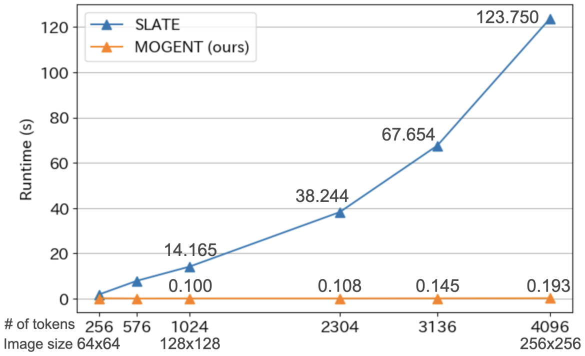


Figure 2: Runtime comparison of image generation between SLATE and MOGENT. All results were computed on a single NVIDIA Tesla V100 GPU.

using cosine similarity as the distance metric. (3) To generate a new image, pick concepts from the library and randomly select a slot per concept, and decode using MOGENT and DVAE decoder. Implementation-wise, SlotDiffusion (Wu et al., 2023b)[1] proposes a simplified version of the evaluation in which they generate new images by randomly shuffling the extracted slots within a batch. As they report that the FID result is close to the aforementioned method, we use their implementation to evaluate the performance on this task.

4 Experiments

We evaluate the benefits of MOGENT over using an autoregressive transformer decoder in terms of (1) computational efficiency, (2) image segmentation ability, and (3) compositional generation ability. We select SLATE (Singh et al., 2022a) as the baseline, as it uses the same transformer-based decoder but with trained on next token prediction. We evaluate on two datasets with distinct characteristics: the 3D Shapes dataset (Burgess & Kim, 2018) and the CLEVR dataset (Johnson et al., 2017). 3D Shapes dataset consists of 400K training images of 3D objects procedurally generated from 6 ground truth independent latent factors, such as color, size, and shape. CLEVR dataset consists of 200K images of multiple objects with random colors and shapes under photorealistic lighting conditions. The images are size 64×64 and 128×128, respectively. We set the number of iteration steps for decoding to $T = 20$ except for its ablational study in Section 4.4. Hyperparameters and training details are summarized in Appendix A. Experiment setups for image segmentation and compositional editing tasks are summarized in Section A.2.

4.1 Computation Efficiency

We evaluated the computational efficiency of MOGENT against SLATE, focusing on training and inference speed. We report the number of parameters, time per training step, and the time required to generate a single image (Table 1). All metrics were measured on a single NVIDIA Tesla V100 GPU. As the table shows, MOGENT has marginally more parameters compared to SLATE, primarily due to an additional linear layer used to project the outputs of the masked transformer decoder into token probabilities. However, MOGENT speeds up training and generation speed around 10 times faster by leveraging the parallel decoding scheme enabled by the masked bidirectional transformer.

To further investigate the efficiency advantage, we compared the image generation runtime of MOGENT and SLATE across varying image resolutions. As illustrated in Figure 2, the relative speedup offered by MOGENT becomes even more pronounced as image resolution increases.

---

[1] `https://github.com/Wuziyi616/SlotDiffusion`.

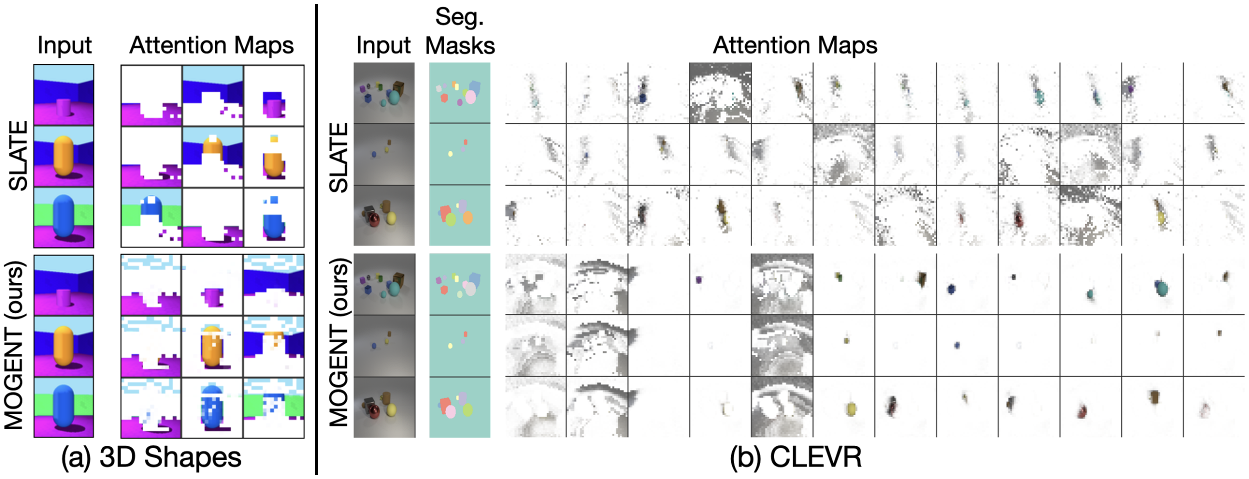


Figure 3: Visualization of attention maps of SLATE and MOGENT on (a) 3D Shapes and (b) CLEVR with masks dataset. For CLEVR, we plot the ground-truth segmentation masks in the second column.

Table 2: Comparison of SLATE and MOGENT on the image segmentation task on CLEVR with masks dataset. We report FG-ARI, mIoU, FG-mIoU, and mBO.

	FG-ARI (↑)	mIoU (↑)	FG-mIoU (↑)	mBO (↑)
SLATE	0.566	0.253	0.233	0.242
MOGENT (Ours)	**0.852**	**0.576**	**0.581**	**0.595**

4.2 Image Segmentation

We evaluate how well the models disentangle the images into individual objects. To assess this, we evaluate the image segmentation performance using the CLEVR dataset with ground-truth segmentation masks provided by Greff et al. (2019). We report four metrics; (1) foreground Adjusted Rand Index (FG-ARI), (2) mean Intersection over Union (mIoU), (3) foreground mIoU (FG-mIoU), and (4) mean Best Overlap (mBO). These metrics quantify how well the predicted object masks match the ground-truth segmentation. Table 2 presents the segmentation ability of the models. As the table shows, MOGENT outperforms SLATE across all metrics. As shown in Figure 3, MOGENT has learned to disentangle the images into objects better than the baseline model across the two datasets.

4.3 Compositional Generation

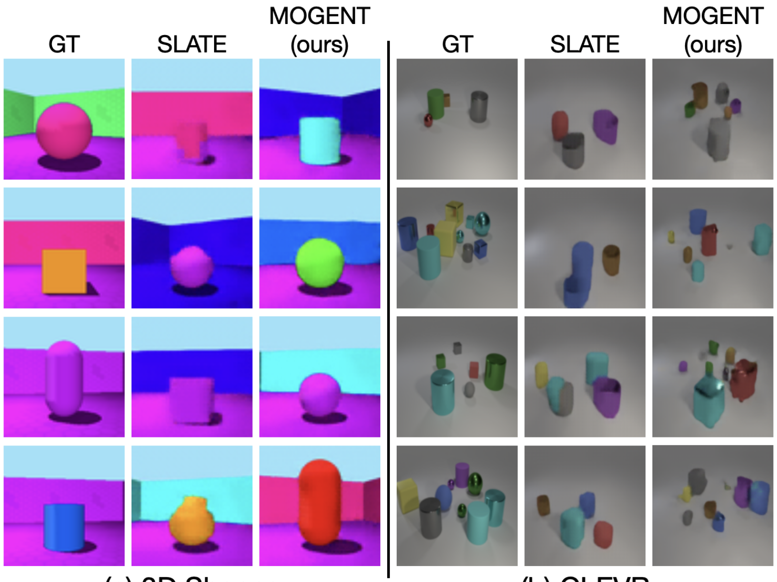
In this section, we evaluate how well the model is able to generate novel scenes by combining the learned representations. We conduct two experiments; (1) compositional generation task described in Section 3.2 and (2) compositional editing task.

Compositional generation. We assess the Fréchet Inception Distance (FID) (Heusel et al., 2017) score and Inception Score (IS) (Salimans et al., 2016) on the compositional generation task in Table 3. We calculate FID score and IS between 40K generated images and the ground-truth images. Figure 4 shows examples of the generated images on both datasets. MOGENT achieves better FID score on both datasets, with larger improvements on the CLEVR dataset. In terms of IS, our model shows better score on the 3D Shapes dataset and competitive one for the CLEVR dataset. Combined with the qualitative results, it shows that MOGENT is able to reuse the learned representations to generate new scenes.

Compositional editing. In Figure 5, we apply MOGENT to image editing using the learned slots. Using the CLEVR dataset, we randomly swap an inferred slot representation between two images. We swap the

Table 3: Comparison of SLATE and MOGENT on the compositional generation task measured by FID score and IS. We report reproduced results for SLATE.

Dataset	FID (↓)		IS (↑)	
	SLATE	MOGENT (Ours)	SLATE	MOGENT (Ours)
3D Shapes	46.51	**44.96**	3.35	**3.70**
CLEVR	116.47	**72.23**	**2.75**	2.46



(a) 3D Shapes (b) CLEVR

Figure 4: Visualization of compositional generations results on (a) 3D Shapes and (b) CLEVR dataset. Across both datasets, MOGENT is able to generate more realistic and diverse images compared to SLATE.

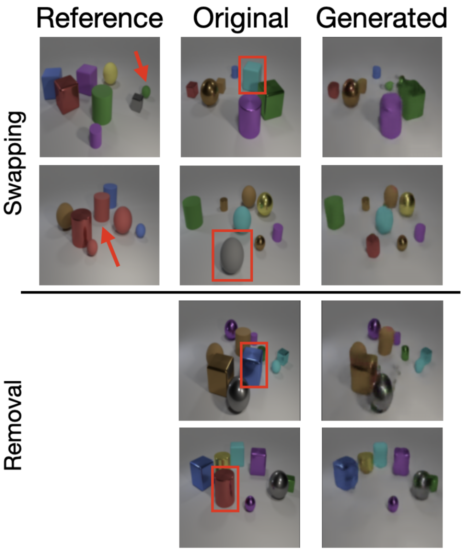


Figure 5: Visualization of compositional editing on the CLEVR dataset. Red squares represent the target object we aim to swap or remove and red arrows represent with which object we swap the target object with.

slot with either slot of the foreground objects or the background to conduct object swapping or removal, respectively. We use the attention map from QSA slot encoder to mask the tokens where the attention values are high. Then, we generate the image, similar to image inpainting task. While autoregressive models require the entire image to be generated from scratch, MOGENT can easily edit parts of the image as it does not have any restrictions regarding token prediction orders. As the figure shows, MOGENT is able to swap and insert an object from a different scene. We can also remove objects by swapping the slot with a slot representing the background. Our model generates realistic images that retains the objects relationships such as appearance and occlusion.

4.4 Ablations

Model Design. We conduct an ablation study evaluating three architectural and training design choices of our model explained in Section 3.2 (Table 4). Additionally, we experiment using rotary positional embeddings (RoPE) (Su et al., 2024) in the transformer decoder (Table 5).

As Table 4 shows, adopting QSA (Jia et al., 2023) over standard Slot Attention slightly improves both FID and IS metrics. Moreover, further adding zero mask init on the `[MASK]` token and training the model using cross-entropy loss on only the masked tokens both contribute positively to the model's generation ability, leading to largely improved FID scores. We also visualize extracted slots on the 3D Shapes dataset using t-SNE (Van der Maaten & Hinton, 2008) and codebook usage of the transformer decoder in Figure 10

Table 4: Ablation on using QSA, zero mask init, and calculating loss on only masked tokens (mask loss). We report FID score and IS on the compositional generation task using 3D Shapes dataset.

QSA	zero mask init	mask loss	FID (↓)	IS (↑)
✗	✗	✗	136.36	3.61
✓	✗	✗	130.69	**3.96**
✓	✓	✗	55.87	3.79
✓	✓	✓	**44.96**	3.70

Table 5: Ablation of using RoPE. We report FID score and IS on the compositional generation task for 3D Shapes and CLEVR datasets.

Dataset	RoPE	FID (↓)	IS (↑)
3D Shapes	✗	**44.96**	**3.70**
	✓	52.79	3.60
CLEVR	✗	91.35	2.33
	✓	**72.23**	**2.46**

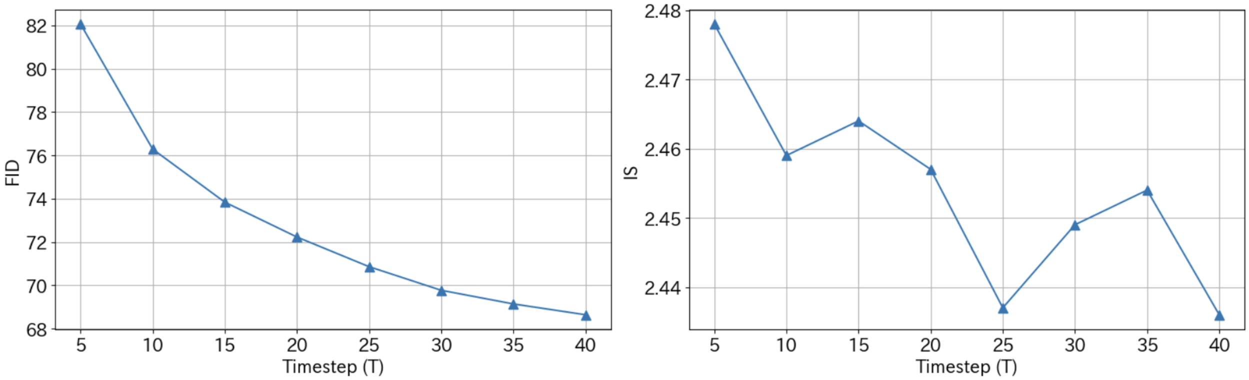


Figure 6: Ablation on the number of iteration steps measured on the compositional generation task using the CLEVR dataset. We report FID score (left) and IS (right).

and Figure 11, respectively. As the figure shows, adding zero mask init and mask loss contributes largely in improving model's compositional generation task by enabling better slot disentanglement and avoiding codebook collapse, respectively. These findings show that integrating components in both the slot encoder and decoder is important for achieving effective object-centric representation learning with MOGENT.

Table 5 shows the effect of using RoPE as positional embeddings in the transformer decoder. As the table shows, we see that RoPE is effective particularly when training the model on the CLEVR dataset. We hypothesize that this improvement stems from the characteristics of CLEVR, which features images with more variation in object size, including smaller objects. In such scenarios, adding RoPE allows MOGENT to better utilize the relative distance between tokens to capture local details and distinguish individual objects. Consequently, RoPE reinforces the spatial locality important for effective object-centric learning and helps prevent over-reliance on the global context alone.

Iteration Number. We study the effect of the number of iterations (T) on our model by evaluating compositional generation task with different Ts. As shown in Figure 6, increasing T does not necessarily yield consistent improvements: while higher T leads to improved FID scores, it simultaneously results in decreased IS. This observation differs from previous findings in masked image modeling (Chang et al., 2022a), where both FID and IS initially improved with increasing iterations until a "sweet spot", beyond which performance declined. We hypothesize that, since slots encode both object identity and positional information, it acts as a strong conditioning during generation compared to other conditionings, such as text, label and layout. Therefore, we think that increasing iterations does not enhance the diversity of generated images.

5 Conclusion

In this work, we addressed the computational challenge posed by prevalent object-centric generative models. While previous works offer powerful capabilities for compositional scene generation from object-centric representation, their model architecture makes the model computationally inefficient, especially during generation. We proposed MOGENT, an object-centric representation learning architecture using a masked generative modeling approach, inspired by MaskGIT. Using the iterative decoding scheme, MOGENT is able to decode slots to tokens in parallel, achieving efficient decoding regardless of input size.

Our results demonstrate that MOGENT reduces computational requirements up to 10x speedup in generation compared to autoregressive baselines. As the number of steps to generate is independent from the image resolution, the relative speedup by our model increases as the input size increases. Importantly, efficiency is achieved while maintaining to learn to generate images from object-centric representations, achieving strong performance on object segmentation and compositional generation tasks on 3D Shapes and CLEVR datasets. Furthermore, our ablation studies validate that incorporating appropriate inductive biases, such as using QSA and initializing mask embeddings as zeros, is crucial for effective masked generative modeling and object-centric representation learning. We empirically show that these improvements add the spatial locality bias and avoid codebook collapse of the decoder, both needed for achieving meaningful disentanglement. Overall, our work establishes masked generative modeling as a viable and highly efficient alternative for object-centric generation.

Despite its success, MOGENT has limitations. Firstly, our current work primarily uses synthetic or semi-synthetic datasets where objects are well-defined and separated. In natural scenes, the definition of an "object" becomes more ambiguous; boundaries are often unclear, objects exhibit complex articulation or occlude significantly, and distinguishing foreground objects from the background is non-trivial. Secondly, while significantly faster, the masked generative approach of MOGENT differs from the iterative refinement process of diffusion models. While diffusion models can refine the entire generation throughout their sampling process, masked generative models do not have the mechanism to correct its previously sampled tokens. This lack of a refinement mechanism can potentially lead to uncorrected errors or visual artifacts appearing in the final output (Lezama et al., 2022).

Future work should focus on addressing these limitations. This includes scaling MOGENT to more complex, real-world image and video datasets. Following prior works (Seitzer et al., 2023; Wu et al., 2023b; Zadaianchuk et al., 2023), further investigation of employing pretrained visual transformers (Dosovitskiy et al., 2020) such as DINO (Caron et al., 2021) or DINOv2 (Oquab et al., 2024) to extract patch-level representations from scenes could yield additional insights. Investigation on hybrid approaches such as masked autoregressive models (Li et al., 2024) could also be helpful in solving the inability of masked generative modeling to refine during generation. In addition, we leave the application of MOGENT to downstream tasks such as reinforcement learning and reasoning for future exploration.

Broader Impact Statement

Our framework proposes an efficient, object-centric representation learning on various datasets. Our model shows promising direction towards in using object-centric representations for efficient generation or editing tasks. Since we mainly experiment on synthetic or semi-synthetic datasets, we do not see any immediate risks of human rights violations or security threats in our work. However, future works should investigate on the scalability of our work and evaluate the potential risks.

References

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

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

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Ayush Chakravarthy, Trang Nguyen, Anirudh Goyal, Yoshua Bengio, and Michael C Mozer. Spotlight attention: Robust object-centric learning with a spatial locality prior. *arXiv preprint arXiv:2305.19550*, 2023.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022a.

Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4055–4075. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/chang23b.html`.

Michael Chang, Tom Griffiths, and Sergey Levine. Object representations as fixed points: Training iterative refinement algorithms with implicit differentiation. *Advances in Neural Information Processing Systems*, 35:32694–32708, 2022b.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, June 2019.

Aniket Rajiv Didolkar, Andrii Zadaianchuk, Anirudh Goyal, Michael Curtis Mozer, Yoshua Bengio, Georg Martius, and Maximilian Seitzer. On the transfer of object-centric representation learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=bSq0XGS3kW`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

SM Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Geoffrey E Hinton, et al. Attend, infer, repeat: Fast scene understanding with generative models. *Advances in neural information processing systems*, 29, 2016.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6112–6121, 2019.

Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266):20210068, 2022.

Anirudh Goyal, Alex Lamb, Phanideep Gampa, Philippe Beaudoin, Charles Blundell, Sergey Levine, Yoshua Bengio, and Michael Curtis Mozer. Factorizing declarative and procedural knowledge in structured, dynamical environments. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=VVdmjgu7pKM`.

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. Neural expectation maximization. *Advances in neural information processing systems*, 30, 2017.

Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *International conference on machine learning*, pp. 2424–2433. PMLR, 2019.

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1900–1910, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. *Advances in neural information processing systems*, 31, 2018.

Daniel Jiwoong Im, Sungjin Ahn, Roland Memisevic, and Yoshua Bengio. Denoising criterion for variational auto-encoding framework. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=rkE3y85ee`.

Baoxiong Jia, Yu Liu, and Siyuan Huang. Improving object-centric learning with query optimization. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=_-FN9mJsgg`.

Jindong Jiang, Sepehr Janghorbani, Gerard De Melo, and Sungjin Ahn. Scalor: Generative world models with scalable object representations. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SJxrKgStDH`.

Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=gbOukzirpK`.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Ioannis Kakogeorgiou, Spyros Gidaris, Konstantinos Karantzalos, and Nikos Komodakis. Spot: Self-training with patch-order permutation for object-centric learning with autoregressive transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22776–22786, 2024.

Sungwoong Kim, Daejin Jo, Donghoon Lee, and Jongmin Kim. Magvlt: Masked generative vision-and-language transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23338–23348, 2023.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Thomas Kipf, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=aD7uesX1GF_`.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Draft-and-revise: Effective image generation with contextual rq-transformer. *Advances in Neural Information Processing Systems*, 35: 30127–30138, 2022.

José Lezama, Huiwen Chang, Lu Jiang, and Irfan Essa. Improved masked image generation with token-critic. In *European Conference on Computer Vision*, pp. 70–86. Springer, 2022.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2024.

Zhixuan Lin, Yi-Fu Wu, Skand Peri, Bofeng Fu, Jindong Jiang, and Sungjin Ahn. Improving generative imagination in object-centric world models. In *International conference on machine learning*, pp. 6140–6149. PMLR, 2020a.

Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. Space: Unsupervised object-oriented scene representation via spatial attention and decomposition. In *International Conference on Learning Representations*, 2020b. URL `https://openreview.net/forum?id=rkl03ySYDH`.

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.

Elman Mansimov, Alex Wang, Sean Welleck, and Kyunghyun Cho. A generalized framework of sequence generation with application to undirected sequence models. *arXiv preprint arXiv:1905.12790*, 2019.

David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36, 2024.

Akihiro Nakano, Masahiro Suzuki, and Yutaka Matsuo. Interaction-based disentanglement of entities for object-centric world models. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=JQc2VowqCzz`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=a68SUt6zFt`. Featured Certification.

Suraj Patil, William Berman, Robin Rombach, and Patrick von Platen. amused: An open muse reproduction. *arXiv preprint arXiv:2401.01808*, 2024.

Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. Bamm: bidirectional autoregressive motion model. In *European Conference on Computer Vision*, pp. 172–190. Springer, 2024a.

Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1546–1555, 2024b.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Mehdi SM Sajjadi, Daniel Duckworth, Aravindh Mahendran, Sjoerd Van Steenkiste, Filip Pavetic, Mario Lucic, Leonidas J Guibas, Klaus Greff, and Thomas Kipf. Object scene representation transformer. *Advances in neural information processing systems*, 35:9512–9524, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. In *International Conference on Learning Representations*, 2023.

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

Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. *Advances in Neural Information Processing Systems*, 35:18181–18196, 2022b.

Gautam Singh, Yeongbin Kim, and Sungjin Ahn. Neural systematic binder. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=ZPHE4fht19t`.

Elizabeth S Spelke. Principles of object perception. *Cognitive science*, 14(1):29–56, 1990.

Elizabeth S Spelke. Where perceiving ends and thinking begins: The apprehension of objects in infancy. In *Perceptual development in infancy*, pp. 197–234. Psychology Press, 2013.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Sjoerd van Steenkiste, Michael Chang, Klaus Greff, and Jürgen Schmidhuber. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=ryH20GbRW`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=vOEXS39nOF`.

Nicholas Watters, Loic Matthey, Christopher P Burgess, and Alexander Lerchner. Spatial broadcast decoder: A simple architecture for learning disentangled representations in vaes. *arXiv preprint arXiv:1901.07017*, 2019.

Yi-Fu Wu, Minseung Lee, and Sungjin Ahn. Neural language of thought models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=HYyRwm367m`.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models. In *The Eleventh International Conference on Learning Representations*, 2023a.

Ziyi Wu, Jingyu Hu, Wuyue Lu, Igor Gilitschenski, and Animesh Garg. Slotdiffusion: Object-centric generative modeling with diffusion models. *Advances in Neural Information Processing Systems*, 36:50932–50958, 2023b.

Wilson Yan, Danijar Hafner, Stephen James, and Pieter Abbeel. Temporally consistent transformers for video generation, 2023.

Yafei Yang and Bo Yang. Promising or elusive? unsupervised object segmentation from real-world single images. *Advances in Neural Information Processing Systems*, 35:4722–4735, 2022.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10459–10469, 2023.

Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A Ross, and Lu Jiang. Language model beats diffusion - tokenizer is key to visual generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=gzqrANCF4g`.

Andrii Zadaianchuk, Maximilian Seitzer, and Georg Martius. Object-centric learning for real-world videos by predicting temporal feature similarities. *Advances in Neural Information Processing Systems*, 36, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

A Additional Implementation Details

A.1 Hyperparameters and Training Details

The hyperparameters used for our experiments are reported in Table 6. We followed the implementation of SLATE (Singh et al., 2022a) and mainly changed only the transformer decoder architecture. Although MaskGIT (Chang et al., 2022a) uses a larger transformer decoder, with 24 layers, 8 attention heads, 768 embedding dimensions and 3072 hidden dimensions, we kept our hyperparameters similar to the transformer decoder used by SLATE to measure performance fairly. The model was trained using Adam optimizer (Kingma, 2014) with $\beta_1 = 0.9, \beta_2 = 0.999$. We used a fixed learning rate of 3e-4 for the DVAE and a learning rate of 1e-4 with linear warmup for stable learning. For training on 3D Shapes dataset, we halved the learning rate if the validation loss did not decrease for 4 consecutive epochs.

Following QSA (Jia et al., 2023), we added a perturbation to the initial slots by sampling from a normal distribution of mean zero and variance σ for better performance. We applied cosine annealing to decrease the perturbation from 1 to 0 during training. We set the # of annealing steps to match the # of annealing steps for DVAE's temperature. Following SlotDiffusion (Wu et al., 2023b), we set the # of warmup and annealing steps to match 5% and 15% of the total training steps, respectively. During training, we mask the tokens based on a cosine scheduling: for each training sample, the masking rate is sampled from a trncated arccos distribution with density function, $p(r) = \frac{2}{\pi}\left(1 - r^2\right)^{-\frac{1}{2}}$. This has an expected masking rate of 0.64, showing a bias towards higher masking rate. To train MOGENT, while some works (Wu et al., 2023b; Jiang et al., 2023) that pretraining DVAE leads to better performance, we found that training all components from scratch led to better object-centric disentanglement.

During training, MOGENT requires approximately 45GB of memory, which is equivalent to the requirement of SLATE. Training MOGENT takes around 7 days on a single NVIDIA RTX A6000 GPU, while SLATE is trained in 3 days using the same GPU setup. We find that MOGENT requires around twice the number of training steps for training, as masked token prediction is more difficult than next token prediction. We report # of hyperparameters and training and generation speed in Section 4.1.

We reproduced the results for the baseline model, SLATE, as only the code on 3D Shapes dataset was available. To train SLATE, we used the hyperparameters that was reported in the original paper.

Table 6: Hyperparameters of MOGENT.

Dataset		3D Shapes	CLEVR
Batch Size		50	64
Epochs		80	400
Learning Rate Warmup Steps		30000	21860
Max Learning Rate		1e-4	1e-4
Gradient Clipping		1.0	1.0
Encoder	Image Size	64	128
	# of Tokens	256	1024
DVAE	Vocabulary Size	1024	4096
	Max Temperature	1.0	1.0
	Min Temperature	0.1	0.1
	Temp. Annealing Steps	30000	65580
	Learning Rate (w/o warmup)	3e-4	3e-4
Slot Attention	# of Slots	3	12
	# of Iterations	3	3
	Slot Dimension	192	192
	MLP Dimension	192	384
	σ Annealing Steps	30000	65580
MOGENT	# of Layers	4	8
	# of Heads	8	8
	Embedding Dimension	192	192
	Hidden Dimension	192	192

A.2 Experiment Setup

Image Segmentation. As explained in Section 4.2, we use foreground Adjusted Rand Index (FG-ARI), mean Intersection over Union (mIoU), foreground mIoU (FG-mIoU), and mean Best Overlap (mBO) to evaluate segmentation ability. We use the attention map from the Slot Attention encoder and take the argmax along the slot dimension to obtain the predicted mask. To compute mIoU and FG-mIoU, we use Hungarian matching to obtain the ground-truth and slots assignment. To compute mBO, we assign the ground-truth mask to the slot with the largest overlapping mask, and then averages the IoU of all pairs of masks.

Compositional Editing. We view the compositional editing task as an inpainting task. Given a pair of samples, we first extract slot representations. To swap an object, we identify the slots attending to the foreground objects and randomly swap a slot between the samples. To remove an object, we swap with the slot attending to the background.

To generate the edited image, we first identify the tokens corresponding to the slot that was edited by calculating the overlap between the mask per slot and image region per token. We replace the tokens with `[MASK]` token. Finally, we apply the iterative decoding scheme to generate the edited image.

B Additional Experiments

B.1 Image Reconstruction

To assess MOGENT on image reconstruction, we set up by randomly masking the tokens of the input image with a ratio r, and use MOGENT's iterative decoding scheme to reconstruct images. We report two metrics, MSE and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), (Figure 8). We also provide examples of this process in Figure 7.

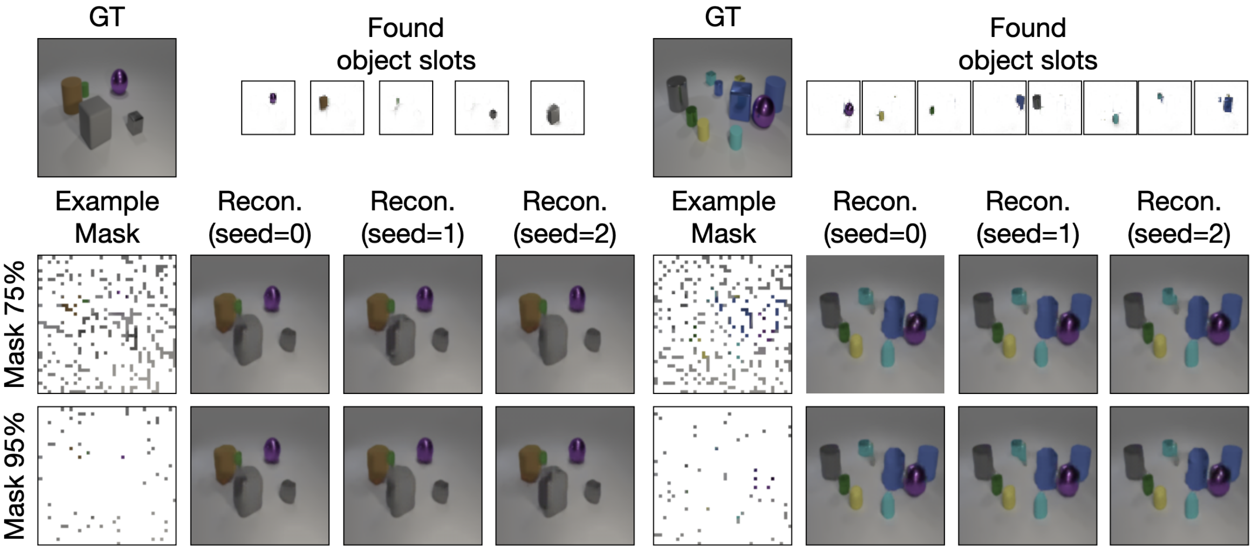


Figure 7: Examples of MOGENT on the image reconstruction task. We show ground-truth image and attention maps of slots for objects in the first row. The second and third row shows example of mask and reconstructed image for $r = 0.75, 0.95$, respectively. On the example on the right, we can see that if MOGENT fails to reconstruct objects if it fails to extract the corresponding slots.

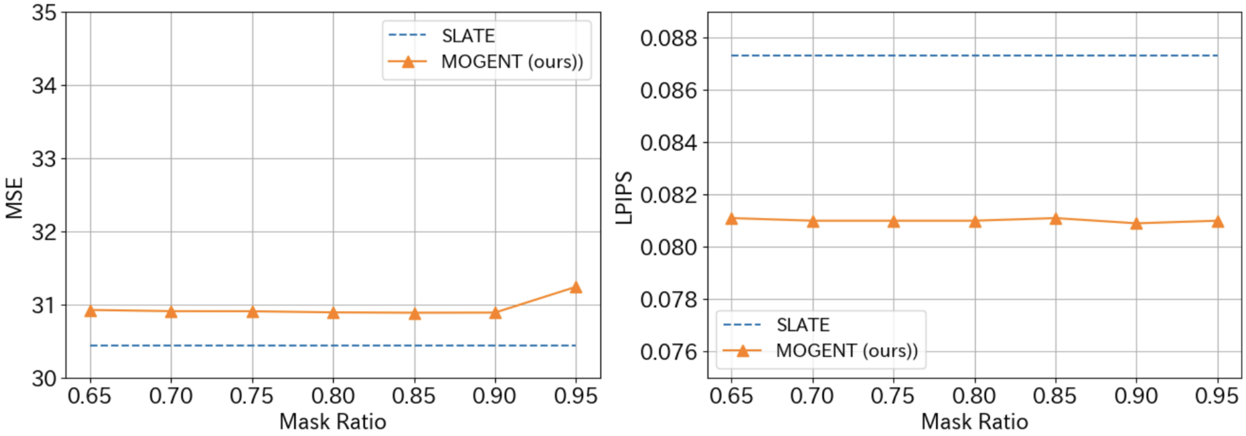


Figure 8: Reconstruction quality vs. mask ratio on CLEVR dataset. We report MSE and LPIPS. We show the score of SLATE in blue as reference.

Interestingly, we find that both MSE and LPIPS do not worsen much even in highly masked setups. Moreover, MOGENT achieves lower LPIPS than SLATE, suggesting that our model reconstructs the images with higher quality than the baseline model. In Figure 7, we compare the reconstruction results between two mask ratio setups. The examples show that even when 95% of the tokens are masked, MOGENT is able to reconstruct the image quite well. We think this is because slot representations contain information about both object identity and position, acting as a strong conditioning about the entire image. In the example on the right, we can see that MOGENT is able to reconstruct all the objects it has found.

B.2 Analysis on Sampling Temperature

The iterative decoding scheme of MOGENT have an option of adding stochasticity by adding noise to the confidence score. Formally, let $\mathbf{s}_t$ be the confidence score of the tokens at iteration t. Then, MOGENT samples

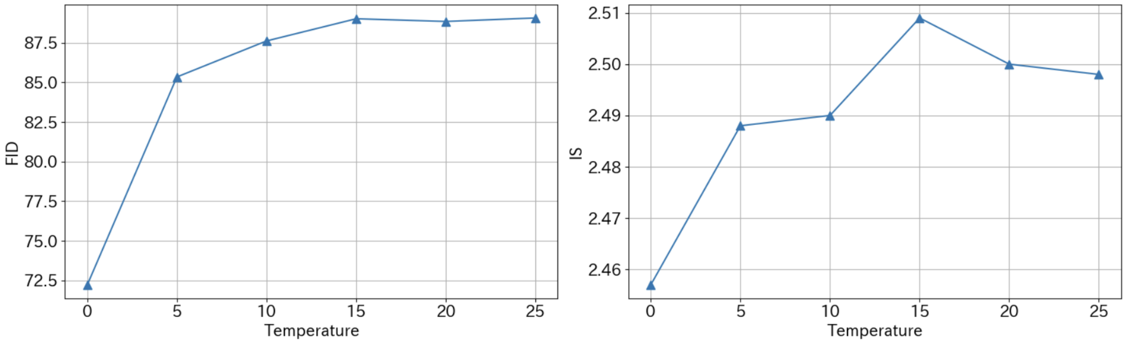


Figure 9: Ablation on the sampling temperature measured on the compositional generation task using the CLEVR dataset. We report FID score (left) and IS (right).

the tokens using $\tilde{\mathbf{s}}_t = \mathbf{s}_t + \tau_{\text{TF}} \cdot (t/T)\mathbf{n}$ as the score, where $\mathbf{n}$ is the sampling noise such as i.i.d. Gumbel noise and τ_{TF} is the sampling temperature.

We assess the effect of sampling temperature using the compositional generation task on the CLEVR dataset (Figure 9). While MaskGIT used a sampling temperature of $\tau_{\text{TF}} = 4.5$, we found that adding stochasticity to the decoding process leads to degradation in performance. We think this is because our model is conditioned on the slots, which contain information about both object identity and position. As this acts as a strong conditioning on the scene appearance and layout, adding noise during sampling to promote diversity leads to worse generation performance.

B.3 Failed Attempts

In this section, we provide records of some model variants we experimented.

Label smoothing. Following MaskGIT, we experimented applying label smoothing when training the masked transformer decoder. However, we found that this often led to unstable training and worse performance on the compositional generation task.

Classifier-free guidance (Ho & Salimans, 2022). We also experimented classifier-free guidance to improve the generation quality of MOGENT. To apply this, we randomly dropped the conditioning (i.e., slots) during training. However, we found that this led to unstable training in which the model did not learn to disentangle the images into individual objects.

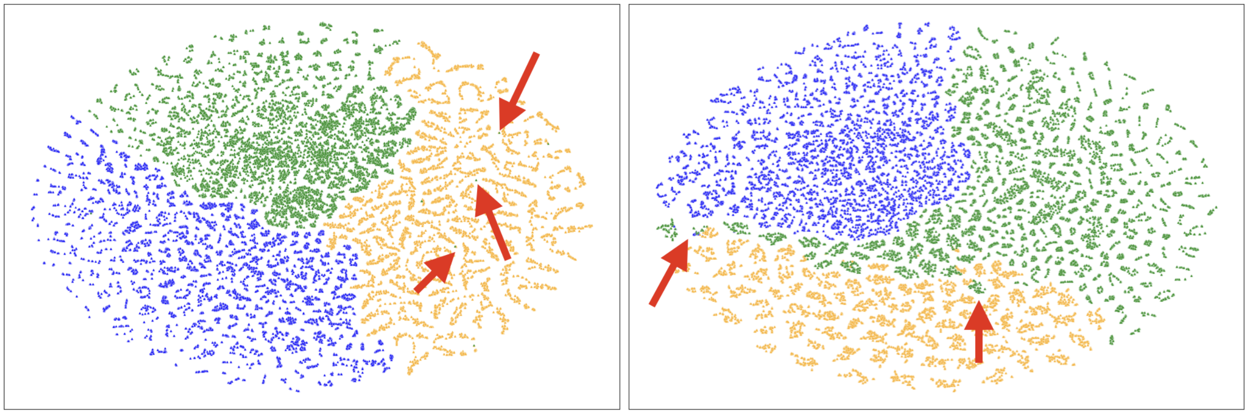
B.4 Further Ablation on Model Design

In thi section, we further analyze the performance gain (Table 4) by adding QSA, zero mask init, and mask loss. First, we visualize the extracted slots on the 3D Shapes dataset using t-SNE (Van der Maaten & Hinton, 2008). We cluster the slots using K-means clustering (Figure 10). As the figure shows, the default model and the model with only QSA have slots that are not disentangled well. By adding zero mask init, we can see that the slots are better disentangled. Next, we plot the codebook usage of the transformer decoder (Figure 11). As the figure shows, between the two models with zero mask init, we see an increase in codebook usage by adding mask loss. These results show that zero mask init and mask loss are especially important as they improve the model's performance by stabilizing training for better slot disentanglement and avoiding codebook collapse, respectively.

C Additional Qualitative Results

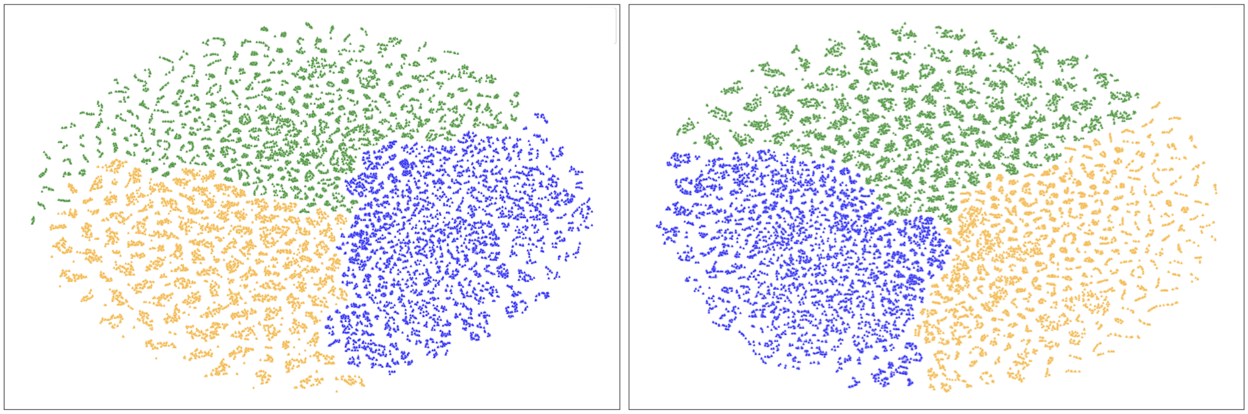
C.1 Image Segmentation

We provide more visualization results of the image segmentation task (Section 4.2) on the CLEVR dataset in Figure 12. Although MOGENT fails to correctly segment objects with smaller size or similar appearances in some cases, our model attends more to individual objects compared to SLATE.



(a) Default

(b) QSA



(c) QSA + zero mask init

(d) QSA + zero mask init + mask loss

Figure 10: t-SNE visualization of extracted slots by different configurations of MOGENT on the 3D Shapes dataset. Colors represent the clustering results using K-means clustering. Red arrows show where some slots are not disentangled well.

C.2 Compositional Geration

Figure 13 and Figure 14 shows more visualizations from the compositional generation task (Section 3.2) on 3D Shapes and CLEVR dataset, respectively. On 3D Shapes dataset, whereas the visual concepts of SLATE look similar between samples (e.g., the attention map for the foreground object), the visual concepts of MOGENT vary more. We think this enabled MOGENT to generate images with higher fidelity. On CLEVR, we see that each visual concept attend to individual objects more, and MOGENT is able to generate images with higher fidelity and diversity.

C.3 Image Reconstruction

Figure 15 shows more visualization results of the image reconstruction task (Section B.1) on the CLEVR dataset. The figure shows that while the reconstruction quality does not change much depending on the mask ratio, the model cannot reconstruct objects which it failed to extract the corresponding slot representations. We can see that these objects tend to be smaller in size, partially occluded by other objects, or have similar colored objects nearby.

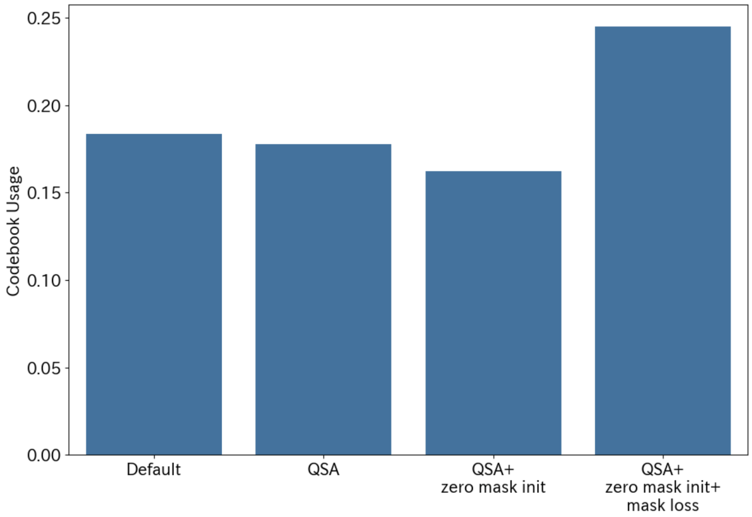


Figure 11: Comparison of the codebook usage of MOGENT's transformer decoder with different configurations.

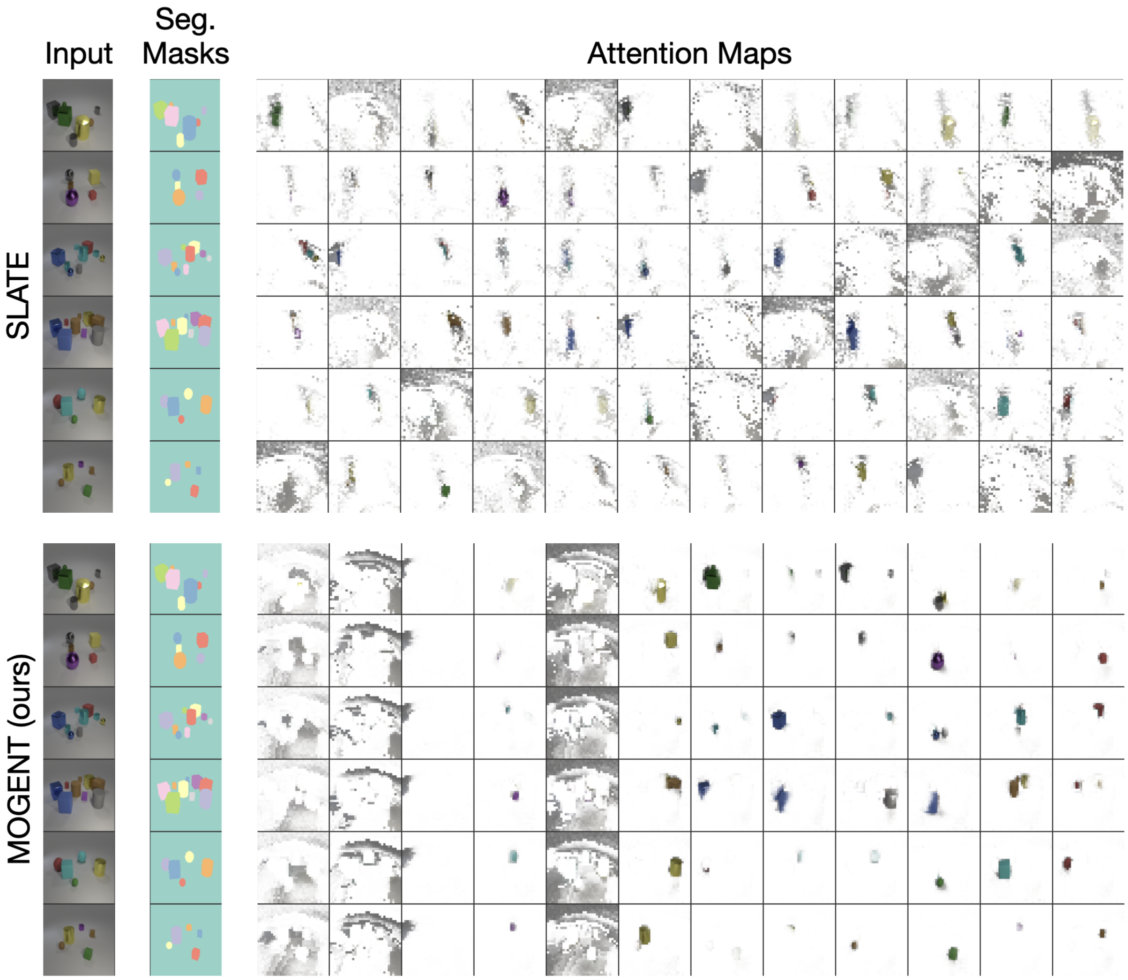


Figure 12: More visualization of attention maps of SLATE and MOGENT on CLEVR with masks dataset.

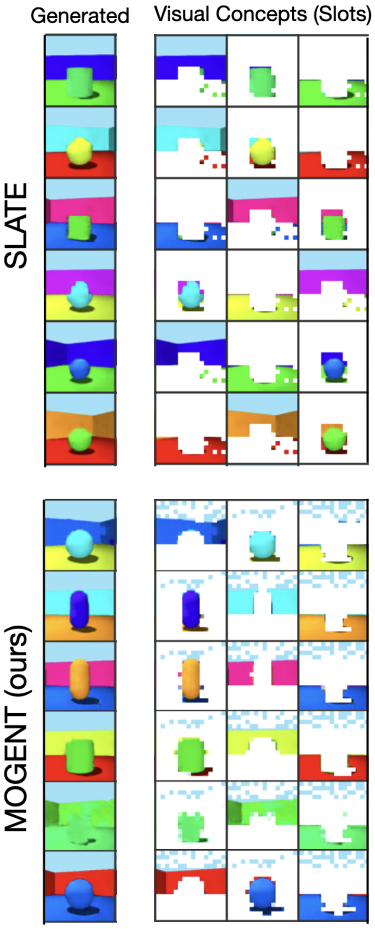


Figure 13: More visualization of generated images by SLATE and MOGENTon the 3D Shapes dataset.

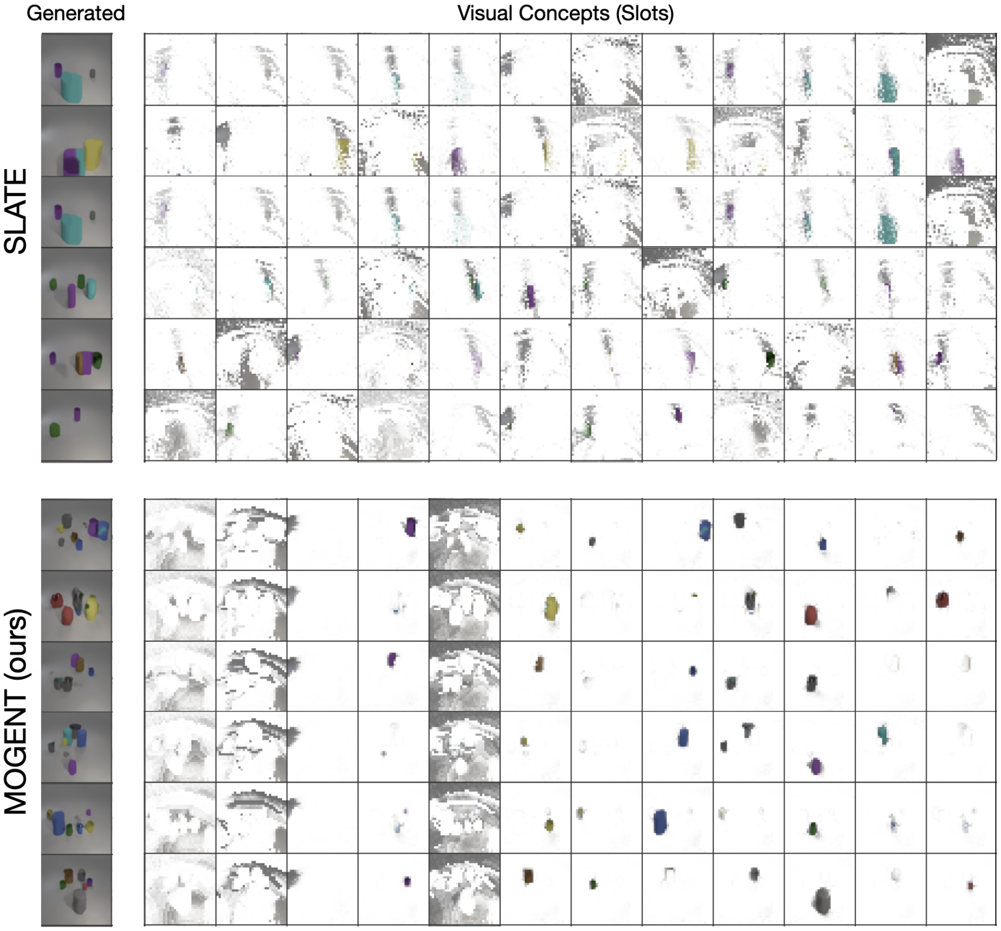


Figure 14: More visualization of generated images by SLATE and MOGENTon the CLEVR dataset.

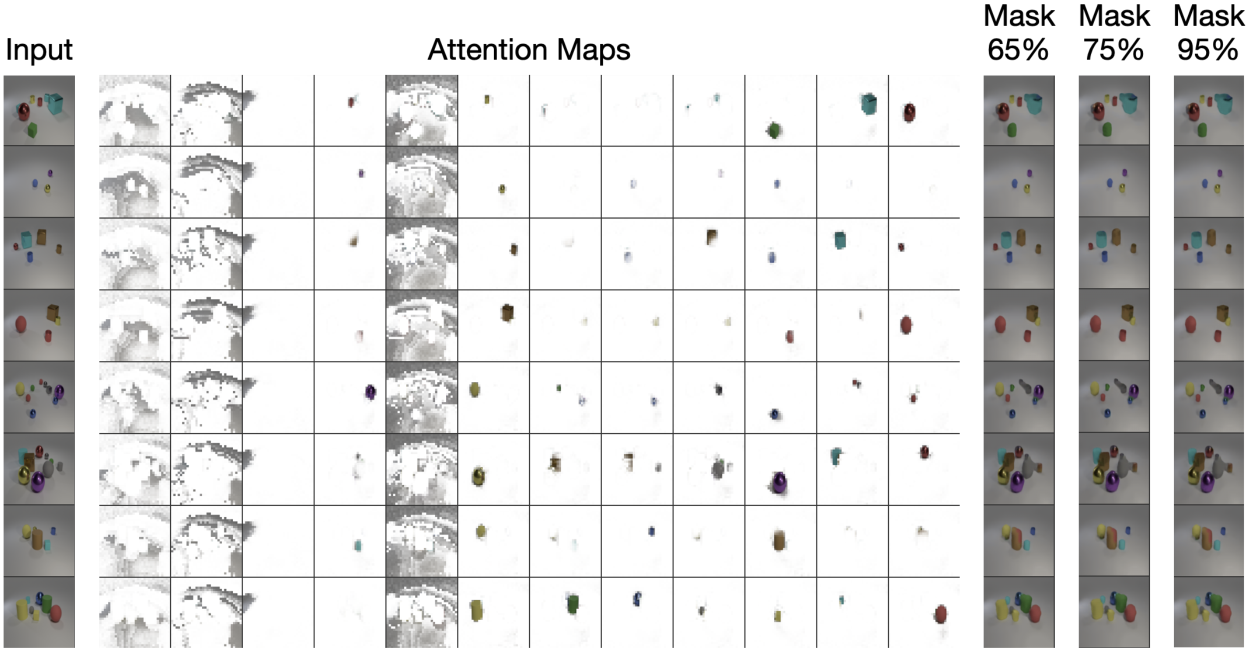


Figure 15: More visualization of attention maps and reconstructed images with different mask ratios on CLEVR with masks dataset.