Object-Centric Slot Diffusion for Unsupervised Compositional Generation

Jindong Jiang[1*] Fei Deng[1] Gautam Singh[1] Sungjin Ahn[2]
[1]Rutgers University [2]KAIST
{jindong.jiang, fei.deng, singh.gautam}@rutgers.edu, sungjin.ahn@kaist.ac.kr

Abstract

Unsupervised compositional generation, a desired ability of object-centric learning, aims to synthesize novel images using visual concepts derived from existing images without supervised guidance. However, existing methods are limited by constraints in image decoders, making them incompetent to handle complex realistic scenes. In this study, we introduce Latent Slot Diffusion (LSD), a novel object-centric learning model that leverages recent advances in diffusion modeling to address these limitations. LSD replaces traditional slot decoders with a slot-conditioned latent diffusion model, resulting in superior performance compared to state-of-the-art approaches in terms of object segmentation and compositional generation. Importantly, for the first time in this line of research, LSD enables unsupervised compositional generation and image editing on the FFHQ dataset. From diffusion models perspective, LSD is the first unsupervised compositional diffusion model that does not rely on supervised annotations, such as text descriptions, for learning to compose.

1. Introduction

The underlying fundamental structure of the physical world is compositional and modular. While in some data modalities like language, this compositional structure is naturally revealed in the form of tokens or words, in general, this structure is hidden in modalities such as images and it is quite elusive how one may discover it. Object-centric learning [18] aims to discover this hidden compositional structure from unstructured observation by learning to bind relevant features into useful tokens unsupervisedly. For images, one of the most popular approaches is to auto-encode the image using a Slot Attention [30] encoder and a *mixture decoder* However, the low-capacity mixture decoder make it struggle when dealing with complex naturalistic scene images. Recently, Singh *et al.* [40] show that increasing the decoder capacity is the key to dealing with complex and naturalistic scenes in object-centric learning. This naturally raises a question: *can the powerful modeling capacity of diffusion models be beneficial for object-centric learning*?

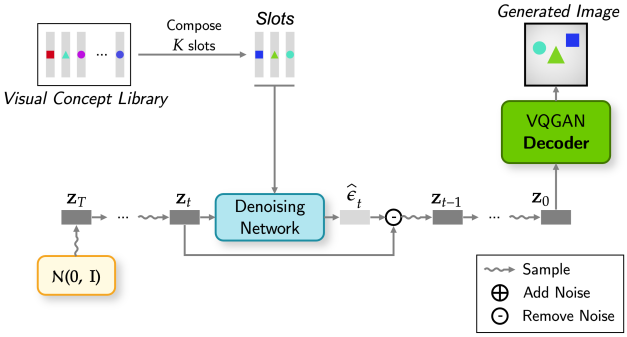

Figure 1. **Compositional Image Generation.** Given the trained model, we can generate novel images by composing a slot-based concept prompt and decoding it using the trained latent slot diffusion decoder.

In this paper, our aim is to answer this question. For this, we propose a novel model called Latent Slot Diffusion (LSD). The LSD model can be understood from two perspectives. From object-centric learning perspective, LSD can be seen as replacing the conventional slot decoders with a conditional latent diffusion model where the conditioning is on object-centric slots provided by Slot Attention. From diffusion models perspective, LSD is the first *unsupervised* compositional conditional diffusion model. Unlike conventional conditional diffusion models [29, 35–37] that rely on supervised annotations, such as text descriptions of an image, to perform compositional generation, LSD constructs such a description with visual concepts learned through unsupervised object-centric learning

In experiments, we show that the LSD model significantly outperform the state-of-the-art model in terms of unsupervised object segmentation and compositional generation.

2. Latent Slot Diffusion

2.1. Object-Centric Encoder

Given an input image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, our object-centric encoder seeks to decompose and represent it as a collection of N vectors or slots $\mathbf{S} \in \mathbb{R}^{N \times D}$ where each slot (denoted

*Correspondence to jindong.jiang@rutgers.edu and sungjin.ahn@kaist.ac.kr.

as $\mathbf{s}_n \in \mathbb{R}^D$) should represent one object in the image. For this, we adopt the Slot Attention architecture.

In Slot Attention, we first encode the input image $\mathbf{x}$ as a set of M input features $\mathbf{E} \in \mathbb{R}^{M \times D_{\text{input}}}$ via a CNN backbone network. Next, we group the features in $\mathbf{E}$ into N *slots* via an iterative refinement procedure. At each step, the slots are refined via *competitive attention* over the input features: $\mathbf{A} = \underset{N}{\texttt{softmax}}\left(\frac{q(\mathbf{S}) \cdot k(\mathbf{E})^T}{\sqrt{D}}\right)$, where, q, k, v are linear projections that map the slots and input features to a common dimension D. Then, for each n, all input features are sum-pooled weighted by their attention weights $\mathbf{A}_{n,m} = \frac{\mathbf{A}_{n,m}}{\sum_{m=1}^{M} \mathbf{A}_{n,m}}$ to produce an attention readout $\mathbf{u}_n = \sum_{m=1}^{M} v(\mathbf{E}_m)\mathbf{A}_{n,m}$. Using the bottom-up information captured by the readout $\mathbf{u}_n$, the slots are updated by an RNN as $\mathbf{s}_n = f_\phi^{\text{RNN}}(\mathbf{s}_n, \mathbf{u}_n)$. In practice, the refinements are performed iteratively several times and slots from the last iteration are considered the final slot representation $\mathbf{S}$.

2.2. Latent Slot Diffusion Decoder

In this section, we describe our proposed decoding approach called *Latent Slot Diffusion Decoder* or *LSD decoder* for reconstructing the image given the slot representation $\mathbf{S}$. Notably, we leverage diffusion modeling to reconstruct the VQGAN latent $\mathbf{z}_0$ conditioned on the slots $\mathbf{S}$. VQGAN provide a way to reduce the computational burden by allowing it to use the lower dimensional latent $\mathbf{z}_0$ as an intermediate reconstruction target. This takes advantage of the recent advances in generative modeling [14, 36].

Sampling Procedure. To sample a $\mathbf{z}_0 \sim p_\theta(\mathbf{z}_0|\mathbf{S})$, we adopt an iterative denoising procedure as in [20, 36]. The sampling process starts with a latent representation $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ filled with random Gaussian noise. Next, conditioned on the slots, we denoise it T times by sampling sequentially from the one-step denoising distribution $\mathbf{z}_{t-1} \sim p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, t, \mathbf{S})$ for $t = T, \ldots, 1$. The one-step denoising distribution is parametrized via a neural network g_θ^{LSD} in the following manner:

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, t, \mathbf{S}) = \mathcal{N}\left(\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{z}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{\epsilon}}_t\right), \beta_t \mathbf{I}\right)$$

Where, $\hat{\boldsymbol{\epsilon}}_t = g_\theta^{\text{LSD}}(\mathbf{z}_t, t, \mathbf{S})$, and $\beta_1, \ldots, \beta_T$ is a linearly increasing variance schedule, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$. The slot conditioning is realized using a cross-attention transformer layer, which calculates the interactions between slots and the denoising feature map. This produces a sequence of latents $\mathbf{z}_T, \mathbf{z}_{T-1}, \ldots, \mathbf{z}_0$ that become progressively cleaner. Finally, $\mathbf{z}_0$ can be considered as the reconstructed latent representation.

Training Procedure. Following LDM [36], the training of $p_\theta(\mathbf{z}_0|\mathbf{S})$ can be cast to a simple procedure for training g_θ^{LSD} as follows. Given an image $\mathbf{x}$, its slot representation $\mathbf{S}$, and its VQGAN latent $\mathbf{z}_0$, we first randomly choose a noise level $t \in \{1, \ldots, T\}$ from a uniform distribution. Given the t, we corrupt the clean latent $\mathbf{z}_0$ and obtain a noised latent $\mathbf{z}_t$ as: $\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t$, where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The noised latent $\mathbf{z}_t$ is then given as input to g_θ^{LSD} along with the slots $\mathbf{S}$ and denoising time-step t to predict the noise $\boldsymbol{\epsilon}_t$. The network g_θ^{LSD} is trained by minimizing the mean squared error between the predicted noise $\hat{\boldsymbol{\epsilon}}_t$ and the true noise $\boldsymbol{\epsilon}_t$: $\mathcal{L}(\phi, \theta) = \mathbb{E}_{t, \mathbf{z}, \boldsymbol{\epsilon}}||\hat{\boldsymbol{\epsilon}}_t - \boldsymbol{\epsilon}_t||^2$

3. Compositional Image Synthesis

In this section, we describe how a trained LSD model can be used to compose and synthesize novel images. Following [40], LSD builds a library of visual concepts from unlabeled images. Then, similarly to composing a sentence prompt using words, we compose a *concept prompt* by picking concepts from this library. We can then synthesize a desired novel image by providing this concept prompt to the LSD decoder.

Unsupervised Visual Concept Library. To build a library of visual concepts from unlabelled images, we first apply slot attention to obtain slots from a large batch of B images. We then collect all these slots as a single set $\mathcal{S}$ and apply K-means on it. We consider the slots that are assigned to a k-th cluster as a visual concept library $\mathcal{V}_k$. This procedure provides K visual concept libraries. Our experiments shall show that this simple K-means procedure can produce semantically meaningful concept libraries. For instance, on a dataset of human face images such as FFHQ [26], the K libraries correspond to useful concept classes such as hair style, face, clothing, and background.

Novel Image Synthesis. Given libraries $\mathcal{V}_1, \ldots, \mathcal{V}_K$, we can compose a concept prompt $\mathbf{S}_{\text{compose}}$ by picking K slots, each from the corresponding k-th library, and stacking them together: $\mathbf{S}_{\text{compose}} = (\mathbf{s}_1, \ldots, \mathbf{s}_K)$, where, $\mathbf{s}_k \sim \text{Uniform}(\mathcal{V}_k)$. We then give the composed prompt $\mathbf{S}_{\text{compose}}$ to the LSD decoder to generate the latent: $\mathbf{z}_{\text{compose}} \sim p_\theta(\mathbf{z}_0|\mathbf{S}_{\text{compose}})$.

4. Related Work

Unsupervised Object-Centric Learning. A common approach of object-centric learning is by auto-encoding. In this line, the focus has been to design an appropriate decoder that supports good decomposition. The most widely used decoders include the mixture-decoder [1, 3, 10–12, 16, 17, 24, 30, 46, 49], spatial transformer decoder [5–7, 13, 23, 27], Neural Radiance Fields (NeRF) [43, 45, 47], transformer decoder [4, 15, 39–42, 48], energy-based models [9] and complex-valued functions [31]. However, existing methods still face limitations in complex scenes or multi-object scenarios.

Diffusion Models. Diffusion models (DMs) are a recent

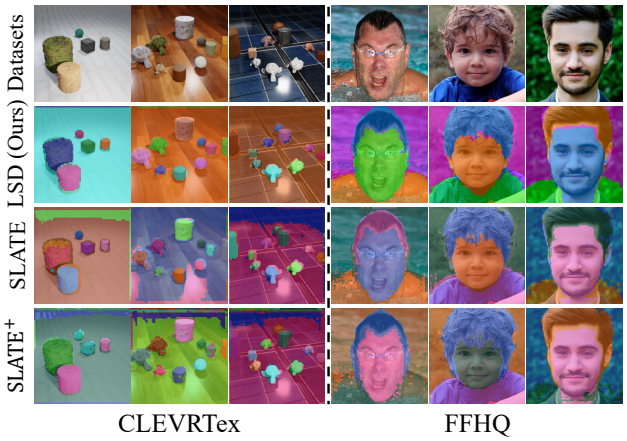

Figure 2. **Visualization of Unsupervised Object Segmentation.** We show visualizations of predicted segments on CLEVRTex and FFHQ datasets.

class of generative models that can produce high-quality images by reversing a stochastic process that gradually adds noise to an image [20, 44]. DMs have been applied to various computer vision tasks [8, 21, 22, 32, 33, 35, 37, 38]. Recently, [36] proposed Latent Diffusion Models (LDM) that operates on a low-dimensional latent space, significantly reducing computational demands. [28] introduced multi-object scene generation, but requires text input. In contrast, our model can generate images compositionally using object concepts directly extracted from images. DMs have also been used for representation learning [2, 34], but these representations are unstructured and not modular like ours.

5. Experiments

We evaluate our proposed Latent Slot Diffusion (LSD) model on unsupervised object segmentation, compositional generation, and image editing. As will be shown, our model significantly outperforms the state-of-the-art on datasets with complex texture and background, including FFHQ [26] which has been beyond the generative capability of object-centric models.

Datasets. We evaluate our model on two datasets including a synthetic multi-object datasets CLEVRTex [25] and FFHQ [26], a dataset of high-quality face images that is beyond the generative capability of current object-centric models. Unlike previous works in this line that only investigate low-resolution images, *e.g.*, 128×128, we use a resolution of 256×256 for all datasets in our experiments.

Baselines. We compare our model against SLATE, the state-of-the-art object-centric learning and unsupervised compositional image generation approach. We use its improved version [42], which is more robust in complex scenes. For a fair comparison with LSD that leverages VQGAN, we also develop a VQGAN-based variant of SLATE denoted as SLATE$^+$, where its low-capacity dVAE [40] is replaced with VQGAN. For all models in this work, we use OpenImages-pretrained VQGAN models [14].

Table 1. **Segmentation and Generation Performance.** We evaluate the segmentation quality using mBO, mIoU and FG-ARI scores and compositional generation quality using the FID score.

(a) Unsupervised Object Segmentation

Segmentation	SLATE	SLATE$^+$	LSD (Ours)
mBO (↑)	51.24	56.22	**66.56**
mIoU (↑)	50.04	54.93	**65.02**
FG-ARI (↑)	43.59	**73.42**	61.74

(b) Compositional Image Generation

FID ↓	SLATE	SLATE$^+$	LSD (Ours)
CLEVRTex	105.83	69.23	**29.53**
FFHQ	112.38	98.76	**27.83**

5.1. Unsupervised Object Segmentation.

In Figure 2, we demonstrate that, without any human annotations, LSD learns to segment the CLEVRTex images into object entities, and the FFHQ images into semantically meaningful components such as face, hair, clothing, and background. We further evaluate the segmentation quality using foreground adjusted rand index (FG-ARI), the mean intersection over union (mIoU), and the mean best overlap (mBO). Our results in Table 1 suggest that LSD significantly outperforms baselines in mBO and mIoU, achieving more than 10% gains in both metrics. We also observe that LSD achieves a lower score on FG-ARI. However, it is important to note that FG-ARI only evaluates the correctness of the foreground pixels and does not account for whether objects are mistakenly considered part of the background or how well the model is able to segment object boundaries, as also highlighted by [12, 25].

5.2. Compositional Generation with Visual Concept Library

Like text-to-image generative models, LSD is able to take unseen slot-based prompts at test time and compose new images. As described in Section 3, we first build a concept library. Then, we sample one slot representation from each visual concept library to form a slot-based prompt. We then feed the slot-based prompts to the diffusion decoder to generate the images. This produces scenes with novel object layouts and faces with unseen attribute combinations.

We report in Table 1b the FID score [19] as a measure of the compositional generation quality. Following standard practice [8], we compute the FID score using 2K generated images and the full training dataset. Across all datasets, LSD achieves significantly better FID scores than SLATE and SLATE$^+$. We further demonstrate the superior compositional generation quality of LSD in Figure 3.

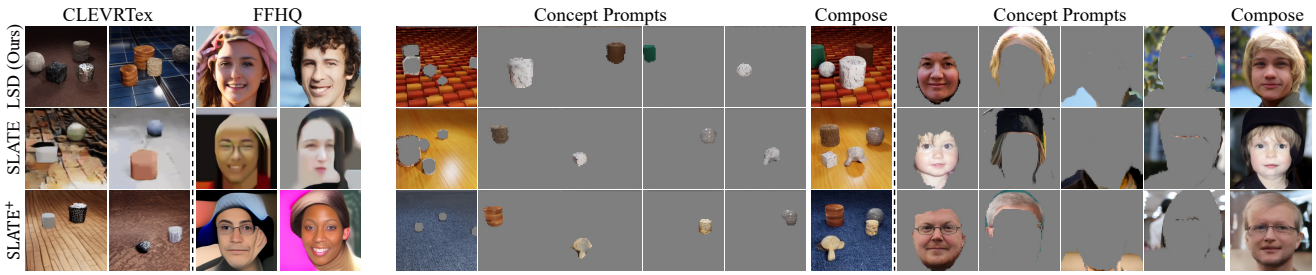

Figure 3. **Compositional Image Generation with Concept Prompts.** *Left:* We show some compositional generation samples. LSD provides significantly higher fidelity and clearer details compared to the other methods. *Right:* We show concept prompts constructed by composing arbitrary slots from our visual concept library and the corresponding generated image by LSD.

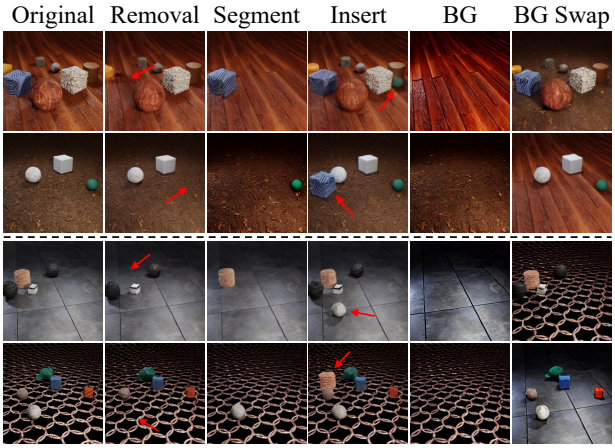

Figure 4. **Slot-Based Object Manipulation.** We show the slot-based image editing ability of our model. In particular, we show edit operations such as object removal, object extraction, object insertion, background extraction, and background swapping.

Figure 5. **Slot-Based Face Replacement.** We show face replacement in the FFHQ dataset, where we compose new images by combining the face slots from Source-B images with the hairstyle, clothing, and background slots from Source-A images.

5.3. Slot-Based Image Editing

In addition to generating new images from randomly sampled slot-based prompts, LSD also allows images editing through slot manipulation. On the CLEVRTex dataset, we perform object removal, single-object extraction, object insertion, background extraction, and background swapping. The results are shown in Figure 4. Object removal and background extraction are achieved by discarding the corresponding slots. Note that, despite the model not encountering single-slot conditioning during training, the background component can be rendered from a single background slot. In the single object extraction task, we render an individual object utilizing the corresponding object slot and the background slot. To demonstrate object insertion and background swapping tasks, we split the image into top and bottom pairs and interchange the corresponding slot before rendering the full scene. We show that the new object or background is rendered in the image coherently.

We further explore face replacement on the FFHQ dataset in Figure 5. LSD decomposes each image into four slots, corresponding to face, hairstyle, clothing, and background. By replacing the face slots of the images, we are able to coherently change the image while maintaining the hairstyle, clothing, and background. The resulting images look realistic, suggesting that LSD can effectively blend various attributes even when given novel combinations.

6. Conclusion

In this work, we proposed the Latent Slot Diffusion model which can be seen in two ways: (1) the first model combining the diffusion models in unsupervised object-centric learning and (2) the first unsupervised compositional diffusion model which does not require supervised annotation like text. We show that the proposed model outperforms the state-of-the-art transformer-based object-centric models in various object-centric tasks. Therefore, we believe that this is a step forward toward object-centric learning that can handle complex naturalistic images, the current main challenge.

References

[1] Titas Anciukevicius, Christoph H Lampert, and Paul Henderson. Object-centric image generation with factored depths, locations, and appearances. *arXiv preprint arXiv:2004.00642*, 2020. 2

[2] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2022. 3

[3] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019. 2

[4] Michael Chang, Alyssa L. Dayan, Franziska Meier, Thomas L. Griffiths, Sergey Levine, and Amy Zhang. Hierarchical abstraction for combinatorial generalization in object rearrangement. In *NeurIPS 2022 Workshop on All Things Attention: Bridging Different Perspectives on Attention*, 2022. 2

[5] Chang Chen, Fei Deng, and Sungjin Ahn. ROOTS: Object-centric representation and rendering of 3D scenes. *Journal of Machine Learning Research*, 22(259):1–36, 2021. 2

[6] Eric Crawford and Joelle Pineau. Spatially invariant unsupervised object detection with convolutional neural networks. In *Proceedings of AAAI*, 2019. 2

[7] Fei Deng, Zhuo Zhi, Donghun Lee, and Sungjin Ahn. Generative scene graph networks. In *International Conference on Learning Representations*, 2020. 2

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3

[9] Yilun Du, Shuang Li, Yash Sharma, Josh Tenenbaum, and Igor Mordatch. Unsupervised learning of compositional energy concepts. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[10] Yilun Du, Kevin Smith, Tomer Ulman, Joshua Tenenbaum, and Jiajun Wu. Unsupervised discovery of 3d physical objects from video. *arXiv preprint arXiv:2007.12348*, 2020. 2

[11] Martin Engelcke, Oiwi Parker Jones, and Ingmar Posner. GENESIS-V2: Inferring unordered object representations without iterative refinement. *arXiv preprint arXiv:2104.09958*, 2021. 2

[12] Martin Engelcke, Adam R Kosiorek, Oiwi Parker Jones, and Ingmar Posner. Genesis: Generative scene inference and sampling with object-centric latent representations. In *International Conference on Learning Representations*. 2, 3

[13] SM Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, and Geoffrey E Hinton. Attend, infer, repeat: Fast scene understanding with generative models. In *Advances in Neural Information Processing Systems*, pages 3225–3233, 2016. 2

[14] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2, 3

[15] Anand Gopalakrishnan, Kazuki Irie, Jürgen Schmidhuber, and Sjoerd van Steenkiste. Unsupervised learning of temporal abstractions with slot-based transformers. *arXiv preprint arXiv:2203.13573*, 2022. 2

[16] Klaus Greff, Raphaël Lopez Kaufmann, Rishab Kabra, Nick Watters, Chris Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. *arXiv preprint arXiv:1903.00450*, 2019. 2

[17] Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Neural expectation maximization. In *Advances in Neural Information Processing Systems*, pages 6691–6701, 2017. 2

[18] Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020. 1

[19] 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. 3

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances In Neural Information Processing Systems*, 2020. 2, 3

[21] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded Diffusion Models for High Fidelity Image Generation. *Journal of Machine Learning Research (JMLR)*, 23:47:1–47:33, 2022. 3

[22] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. *DGMs and Applications @ NeurIPS 2021 Poster*, 2021. 3

[23] Jindong Jiang and Sungjin Ahn. Generative neurosymbolic machines. In *Advances in Neural Information Processing Systems*, 2020. 2

[24] Rishabh Kabra, Daniel Zoran, Goker Erdogan, Loic Matthey, Antonia Creswell, Matthew Botvinick, Alexander Lerchner, and Christopher P Burgess. Simone: View-invariant, temporally-abstracted object representations via unsupervised video decomposition. *arXiv preprint arXiv:2106.03849*, 2021. 2

[25] Laurynas Karazija, Iro Laina, and Christian Rupprecht. Clevrtex: A texture-rich benchmark for unsupervised multi-object segmentation. *arXiv preprint arXiv:2111.10265*, 2021. 3

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 3

[27] 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*, 2020. 2

[28] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua Tenenbaum. Compositional Visual Generation with Composable Diffusion Models. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[29] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 423–439. Springer, 2022. 1

[30] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention, 2020. 1, 2

[31] Sindy Lowe, Phillip Lippe, Maja R. Rudolph, and Max Welling. Complex-valued autoencoders for object discovery. *arXiv preprint arXiv:2204.02075*, 2022. 2

[32] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided Image Synthesis and Editing with Stochastic Differential Equations. *International Conference on Learning Representations*, 2022. 3

[33] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning (ICML)*, pages 16784–16804, 2022. 3

[34] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10619–10629, 2022. 3

[35] A. Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 3

[36] 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*, pages 10684–10695, 2022. 1, 2, 3

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Ayan, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances In Neural Information Processing Systems*, 2022. 1, 3

[38] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image Super-Resolution Via Iterative Refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–14, 2022. 3

[39] Mehdi S. M. Sajjadi, Daniel Duckworth, Aravindh Mahendran, Sjoerd van Steenkiste, Filip Pavetić, Mario Lučić, Leonidas J. Guibas, Klaus Greff, and Thomas Kipf. Object Scene Representation Transformer. *Advances In Neural Information Processing Systems*, 2022. 2

[40] Gautam Singh, Fei Deng, and Sungjin Ahn. Illiterate dall-e learns to compose. In *International Conference on Learning Representations*, 2022. 1, 2, 3

[41] Gautam Singh, Yeongbin Kim, and Sungjin Ahn. Neural Systematic Binder. In *International Conference on Learning Representations*, 2023. 2

[42] Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. *arXiv preprint arXiv:2205.14065*, 2022. 2, 3

[43] Cameron Smith, Hong-Xing Yu, Sergey Zakharov, Frédo Durand, Joshua B. Tenenbaum, Jiajun Wu, and Vincent Sitzmann. Unsupervised discovery and composition of object light fields. *arXiv preprint arXiv:2205.03923*, 2022. 2

[44] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. *International Conference on Learning Representations (ICLR)*, 2021. 3

[45] Karl Stelzner, Kristian Kersting, and Adam R Kosiorek. Decomposing 3d scenes into objects via unsupervised volume segmentation. *arXiv preprint arXiv:2104.01148*, 2021. 2

[46] Julius von Kügelgen, Ivan Ustyuzhaninov, Peter Gehler, Matthias Bethge, and Bernhard Schölkopf. Towards causal generative scene models via competition of experts. *arXiv preprint arXiv:2004.12906*, 2020. 2

[47] Yizhe Wu, Oiwi Parker Jones, and Ingmar Posner. Obpose: Leveraging canonical pose for object-centric scene inference in 3d. *arXiv preprint arXiv:2206.03591*, 2022. 2

[48] Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models. *arXiv preprint arXiv:2210.05861*, 2022. 2

[49] Ruixiang Zhang, Tong Che, B. Ivanovic, Renhao Wang, Marco Pavone, Yoshua Bengio, and Liam Paull. Robust and controllable object-centric learning through energy-based models. *arXiv preprint arXiv:2210.05519*, 2022. 2