SLOT-GUIDED ADAPTATION OF PRE-TRAINED DIFFUSION MODELS FOR OBJECT-CENTRIC LEARNING AND COMPOSITIONAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We present SlotAdapt, an object-centric learning method that combines slot attention with pretrained diffusion models by introducing adapters for slot-based conditioning. Our method preserves the generative power of pretrained diffusion models, while avoiding their text-centric conditioning bias. We also incorporate an additional guidance loss into our architecture to align cross-attention from adapter layers with slot attention. This enhances the alignment of our model with the objects in the input image without using external supervision. Experimental results show that our method outperforms state-of-the-art techniques in object discovery and image generation tasks across multiple datasets, including those with real images. Furthermore, we demonstrate through experiments that our method performs remarkably well on complex real-world images for compositional generation, in contrast to other slot-based generative methods in the literature.

1 INTRODUCTION

The real world is inherently structured with distinct, composable parts and objects that can be combined in various ways; this compositional characteristic is essential for cognitive functions like reasoning, understanding causality, and ability to generalize beyond training data (Lake et al., 2017; Bottou, 2014; Schölkopf et al., 2021; Bahdanau et al., 2019; Fodor & Pylyshyn, 1988). While language clearly reflects this modularity through sentences made up of distinct words and tokens, the compositional structure is less obvious in visual data. Object-centric learning (OCL) offers a promising approach to uncover this latent structure by grouping related features into coherent object representations without supervision (Kahneman et al., 1992; Greff et al., 2020). By decomposing complex scenes into separate objects and their interactions, OCL mimics how humans interpret their environment (Spelke & Kinzler, 2007), potentially improving the robustness and interpretability of AI systems (Lake et al., 2017; Schölkopf et al., 2021). This approach shifts from traditional pixel-based feature extraction to a more meaningful segmentation of visual data, which is key for better generalization and supporting high-level reasoning tasks.

Recent advances in OCL have shown the potential to incorporate powerful generative models, such as transformers and diffusion models, into the OCL framework as image decoders. Notably, models such as Latent Slot Diffusion (LSD) (Jiang et al., 2023) and SlotDiffusion (Wu et al., 2023b) have considerably improved performance in object discovery and visual generation tasks in real-world settings by employing slot-conditioned diffusion models. A concurrent work, GLASS (Singh et al., 2024), has made further progress in handling complex natural scenes by using generated captions as external guiding signals.

There exist two main approaches of integrating diffusion models into slot-based OCL methods in order to deal with real-world images. The first one, as seen in GLASS and LSD, directly uses a pretrained stable diffusion model as decoder. While this approach attempts to fully harness the generative power of pretrained diffusion models, it relies on cross-attention layers to achieve slot conditioning, which were however optimized to receive text input. In contrast, the second approach, as in SlotDiffusion, trains the diffusion model from scratch on the target dataset, thereby eliminating any such text-conditioning related biases. However, this limits the generation capacity of the slot-conditioned model, making it less effective in handling complex real-world images.

Enabling slot-based methods to fully exploit the generative capabilities of pretrained diffusion models while avoiding biases due to text-trained conditioning is a major challenge in object-centric learning, which hinders the effective handling of complex real-world images, particularly in compositional image generation tasks (Jung et al., 2024). In this work, we propose a novel approach that

addresses this challenge by combining slot attention (Locatello et al., 2020) with adaptive conditioning in diffusion models (Mou et al., 2024; Rombach et al., 2022; Sohl-Dickstein et al., 2015; Ho et al., 2020). Our method, that we refer to as SlotAdapt, introduces the use of adapter layers (Mou et al., 2024) for slot conditioning, which eliminates text-trained conditioning bias in pretrained diffusion models and allows the slots to diverge from representations in the text embedding space while retaining the generative power of the pretrained diffusion model.

Another challenge for OCL methods is dealing with the part-whole hierarchy problem – the difficulty of deciding whether to segment an object into its parts or as a whole (Hinton, 1979), especially in real-world settings. The GLASS model (Singh et al., 2024) mitigates this problem with the aid of external supervision and at the cost of increased complexity, whereas Jung et al. (2024) introduce compositional representation learning; however the applicability of their method on real-world datasets remains limited. In order to address this challenge, we propose using cross-attention masks from adapter layers as pseudo-ground truth to guide slot attention maps. This self-supervisory signal enhances the alignment between learned slots and actual objects without external supervision, facilitating the learning of meaningful object representations. In fact, in our architecture, the guidance is mutual, as the cross-attention masks in the diffusion model are also guided and refined jointly with the slot attention masks, improving the image generation process.

In summary, we have the following main contributions: 1) We introduce adapters for slot-based conditioning to combine slot attention with pretrained diffusion models, 2) We propose an additional guidance loss to align cross-attention from adapter layers with slot attention without using external supervision, and 3) We present compositional image generation results on complex real-world images.

Through extensive experiments on various datasets, we demonstrate that our method, SlotAdapt, outperforms previous approaches in object discovery and compositional image generation tasks (Jiang et al., 2023; Wu et al., 2023b; Seitzer et al., 2023), especially on complex real-world image datasets. The performance of SlotAdapt in segmentation task is similar with the concurrent work (Singh et al., 2024), yet we achieve this with significantly reduced computational requirements and without any external supervision. Note that GLASS does not address the compositional generation task. To the best of our knowledge, our work is the first to present successful results for compositional generation on COCO (Lin et al., 2014), a large complex real-world image dataset with 80 different object classes.

2 Related Work

2.1 Unsupervised Object-Centric Learning

Unsupervised object-centric learning aims to decompose visual scenes into meaningful object representations without need for annotation. Early methods used iterative inference and CNN decoders to reconstruct scenes from object-specific feature vectors (*slots*) (Eslami et al., 2016; Burgess et al., 2019; Greff et al., 2019; Lin et al., 2019; Jiang et al., 2019; Lin et al., 2020). Attend-Infer-Repeat (AIR) (Eslami et al., 2016) and Sequential AIR (SQAIR) (Kosiorek et al., 2018) reconstructed objects in canonical poses using patch-based decoders. Slot attention (Locatello et al., 2020) and SAVi (Kipf et al., 2021) employed the spatial broadcast decoder (Watters et al., 2019) to predict images and segmentation masks from slots, combined via alpha compositing. While effective on synthetic datasets like CLEVR (Johnson et al., 2017) and 3D Shapes (Burgess & Kim, 2018), these methods struggle with complex, real-world images. Transformer-based decoders, such as SLATE (Singh et al., 2021) and STEVE (Singh et al., 2022), pre-train a discrete VAE (dVAE) to tokenize images and use slot-conditioned transformers for autoregressive reconstruction. Despite improvements, they face challenges in the high-quality reconstruction of complex real images (Singh et al., 2021; Wu et al., 2023a). Methods like DINOSAUR (Seitzer et al., 2023) bypass reconstruction by using self-supervised learning to discover objects but lack image generation capabilities. A relatively recent research direction in OCL is to employ diffusion models as slot decoders, enhancing scene decomposition and visual fidelity in real-world scenarios (Jiang et al., 2023; Wu et al., 2023b). Several approaches have tackled compositional scene generation through different mechanisms: combining global and symbolic latent variables (Jiang & Ahn, 2020), using hierarchical VAEs with slot attention (Wang et al., 2023), and leveraging hierarchical discrete representations with autoregressive transformers (Wu et al., 2024). While these works focus on synthetic datasets, our work can achieve compositional generation in real-world scenarios.

2.2 Diffusion Models

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) have shown remarkable versatility in tasks such as class-conditioned generation, text-to-image synthesis, and image editing (Dhariwal & Nichol, 2021; Ho & Salimans, 2021; Meng et al., 2021; Ho et al., 2022; Saharia et al., 2022b; Ramesh et al., 2022; Nichol et al., 2022; Saharia et al., 2022a). Latent Diffusion Models (LDMs) (Rombach et al., 2022) address the computational complexity of diffusion models by operating in a lower-dimensional latent space through the use of a pre-trained autoencoder. This approach significantly reduces computational load while maintaining generation quality. LDMs also introduce a flexible conditioning mechanism through cross-attention layers, enabling integration of various conditioning information, such as text embeddings. Recent advances such as T2I adapters (Mou et al., 2024) further enhance the adaptability of LDMs. By introducing additional adapter cross-attention layers, these methods allow fine-tuning for new conditional tasks while keeping the base LDM frozen, reducing computational costs and data requirements for model adaptation.

2.3 Object-Centric Learning Methods with Diffusion Models

Recently, several works have investigated the use of diffusion models in object-centric learning, such as Latent Slot Diffusion (LSD) (Jiang et al., 2023), SlotDiffusion (Wu et al., 2023b), (Jung et al., 2024) and GLASS (Singh et al., 2024). All these methods employ LDMs (Rombach et al., 2022) as slot decoders. Their approaches are similar in that they all use slot attention (Locatello et al., 2020) to extract slot representations from input image and then condition the diffusion model with these slots through cross-attention modules. These methods primarily differ in whether they use pretrained diffusion models, fine-tune them, or train them from scratch. LSD and GLASS use pretrained diffusion models to be able to deal with complex real-world images, assuming an inherent alignment between slot representations and the text-trained cross-attention within the diffusion model, which hinders their performance, especially on image generation tasks. SlotDiffusion, on the other hand, opts to retrain the diffusion model on the target dataset, thereby avoiding biases due to text-trained conditioning, though at the cost of increased computational complexity and, more importantly, at the loss of the generative power of pretrained diffusion models.

GLASS, which is actually a concurrent work (Singh et al., 2024), attempts to mitigate the issues due to text-trained conditioning in pretrained diffusion models by utilizing cross-attention masks as pseudo-ground truth to guide slot attention, in a manner similar to our approach. However, it necessitates an external image captioner or class labels to do this and requires two forward passes: one for generating the training images and pseudo masks, and another for slot attention and image reconstruction. Despite this, GLASS still struggles, particularly in differentiating object instances of the same class, and does not present any compositional image generation results. In contrast, SlotAdapt incorporates additional cross-attention layers into the pretrained diffusion model as adapters, enabling the slots to focus primarily on object semantics, rather than being constrained within a text-centric embedding space.

Jung et al. (2024) have recently introduced a novel framework for compositional learning, which incorporates an additional compositional path into their architecture alongside the slot-conditioned diffusion model. Their approach processes two images simultaneously, composing objects from these images into a single output to apply a diffusion-based generative prior loss. This facilitates the learning of compositional information; however, the applicability of the resulting model on complex real-world images is limited; more specifically, they present results on BDD100K (Yu et al., 2020), an autonomous driving dataset with limited context and relatively small number of object classes.

3 Slot-Based Object-Centric Learning

3.1 Background

Slot Attention: Slot Attention (Locatello et al., 2020) provides a robust framework for segmenting input data into discrete, interpretable components. It operates by dynamically allocating a set of learnable vectors, termed "slots", to represent distinct entities within the input data. These slots are often initialized randomly and then iteratively updated via an attention mechanism, enabling the model to bind each slot to different parts of the input corresponding to individual objects or object-like features. The iterative update rule for each slot is given by

$$\mathbf{U}^{(m)} = \text{Attention}\left(q(\mathbf{S}^{(m)}), k(\mathbf{f}), v(\mathbf{f})\right) \tag{1}$$

$$\mathbf{S}^{(m+1)} = \text{GRU}(\mathbf{S}^{(m)}, \mathbf{U}^{(m)})$$

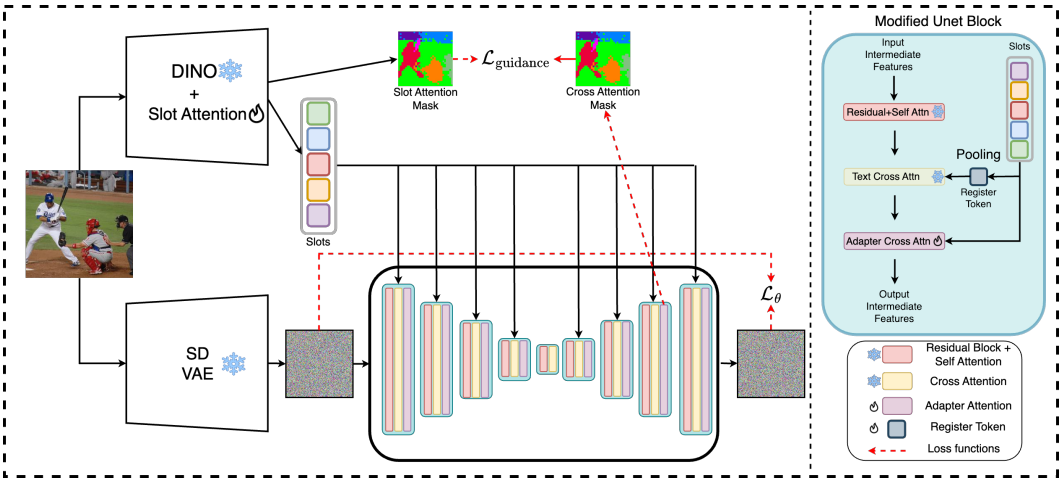

Figure 1: **SlotAdapt Architecture** We extract object-centric information from the input image using a visual backbone, which combines DINO and slot attention. Stable Diffusion VAE is used to encode the image into latent space and then noise is added to the latent. Diffusion process is conditioned on the generated slots as well as the register token which is generated by (mean) pooling the slots. We use the original cross attention layers of diffusion model to condition on the register token, and additional adapter attentions to condition on the learned slots. The overall objective is to predict the noise added to the image. Additionally, we introduce a guidance loss between the slot attention masks and adapter cross attention masks, which encourages the similarity between them. The guidance is only applied in the third upsampling block, while slot conditioning is applied throughout all downsampling and upsampling blocks.

where q, k, v are learnable linear functions, respectively corresponding to the query, key and value in attention computation; $\mathbf{f}$ denotes image features; $\mathbf{S}$ is the set of slots; $\mathbf{U}$ represents the update generated by the attention operation; and m is iteration number for the GRU (Gated Recurrent Unit). The initial slots, $\mathbf{S}^{(0)}$, can be initialized from a Gaussian distribution or using initial object locations (Kipf et al., 2021; Elsayed et al., 2022). Attention is computed over the slot dimension, causing slots to compete for pixels. As a result, each slot attempts to bind to a distinct object in each iteration. The effectiveness of Slot Attention lies in its capacity to disentangle and encode complex scenes into structured representations without supervision. This is achieved through the ability of the mechanism to assign representational capacity where needed.

Diffusion Models: Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) are a powerful class of generative models that simulate the gradual addition of noise to data and then learn to reverse this process. Once trained, they can generate high-quality samples from pure Gaussian noise, and have shown remarkable success in various domains, particularly in computer vision for challenging tasks such as image generation and editing.

The generative process involves a series of reverse diffusion steps from a noise distribution $p(\mathbf{z})$ to the data distribution $p(\mathbf{x})$, defined by conditional probabilities $p(\mathbf{x}_{t-1}|\mathbf{x}_t)$. The training objective minimizes the difference between the actual and generated distributions, typically using a loss function $\mathcal{L}$ that penalizes the expected error of noise prediction at each time step t:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\boldsymbol{\epsilon}_t\sim\mathcal{N}(0,1),t\sim\mathcal{U}(1,T)}\left[\|\boldsymbol{\epsilon}_t - \epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t, t, y)\|_2^2\right], \tag{2}$$

where $\epsilon_{\boldsymbol{\theta}}$ is the noise prediction of the neural network with parameters $\boldsymbol{\theta}$, and y is a conditioning signal for the model such as class or text. When dealing with images, the common practice is to perform training and sampling in a low-dimensional latent space, which is achieved by first transforming the input via a pretrained variational encoder and then to use the standard U-Net architecture as the denoising neural network $\boldsymbol{\epsilon}_\theta$ (Rombach et al., 2022).

Previous OCL methods which employ diffusion models as slot decoders replace the condition y in Eq. 2 with slot representations generated by Slot Attention (Locatello et al., 2020) and use reconstruction loss as a learning signal.

3.2 SLOTADAPT

Object-centric Visual Encoding: We start with an input image $\mathbf{x} \in \mathcal{R}^{H \times W \times 3}$ and transform it through a visual backbone combined with slot attention in a manner similar to recent works (Jiang et al., 2023; Wu et al., 2023b). The visual backbone serves as a feature extractor, reducing the image size to create a set of visual feature vectors $\mathbf{f} \in \mathcal{R}^{h \cdot w \times d}$ by flattening. This reduction makes computations more efficient and condenses key features of the image into a more compact representation. The slot attention mechanism works on these vectors, using a competitive process to generate N slots, represented as $\mathbf{S} \in \mathcal{R}^{N \times d}$. Each slot contains information about a separate object or entity in the scene. The resulting slot representations are then used to condition the diffusion-based decoder as explained in the following section (see also Fig. 1).

Slot Conditional Decoding: We employ a pretrained Stable Diffusion UNet model as a decoder conditioned on the extracted slot representations. This UNet is primarily composed of residual, self-attention, and cross-attention layers stacked together. While self-attention layers capture spatial information, cross-attention layers model semantic relationships between text embeddings and intermediate features. Inspired by recent works (Mou et al., 2024; Ye et al., 2023), we extend this architecture by introducing an adapter layer after each existing cross-attention layer in all downsampling and upsampling blocks of the UNet. These adapter layers are dedicated to conditioning the model on the extracted slots and are essentially cross-attention layers which function similarly to their text-based counterparts in the UNet, with a few minor differences (see Appendix for implementation details).

Our adapter-based conditioning strategy is substantially different from the conventional use of text-trained cross-attention for slot conditioning. Our rationale is that by including these additional cross-attention layers, we enable the slots to focus primarily on object semantics, rather than being constrained within a text-centric embedding space. This is particularly crucial, given that the cross-attention layers in pretrained diffusion modules are typically optimized for text embeddings and hence expect textual input.

We also introduce an extra token that leverages the unused text-conditioning modules of the pretrained UNet to better capture context. This extra token is computed by (mean) pooling either the generated slots or the image features from the visual backbone, and then fed as input into the text cross-attention layers of the UNet model. The main motivation here is to create a 'register' token, similar to the idea presented by Darcet et al. (2024). This register acts as a storage for overall scene information within the diffusion model. By giving this task to a dedicated token, we allow the individual slots to concentrate more on specific objects instead of diluting their focus with background or contextual details.

We freeze the pretrained diffusion model and train only the adapter layers and the slot attention component of our architecture. The training process minimizes a reconstruction objective, formulated as a noise prediction problem. Following (Rombach et al., 2022), at each training iteration (assuming a mini-batch of size 1), time step t, latent image $\mathbf{x}$ and noise $\boldsymbol{\epsilon}_t$ are first sampled from their respective distributions. The noisy image $\mathbf{x}_t$ is then calculated via forward diffusion and fed into the denoising UNet $\epsilon_{\boldsymbol{\theta}}$. The UNet predicts the noise $\boldsymbol{\epsilon}_t$, conditioned on the extracted slots $\mathbf{S}$ and the register token $\mathbf{r}$, and is updated based on the gradient of the following loss function (see also Eq. 2):

$$\mathcal{L}_t(\boldsymbol{\theta}) = \|\boldsymbol{\epsilon}_t - \epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \mathbf{S}, \mathbf{r})\|_2^2 \tag{3}$$

Attention Guidance: The attention mask generated by slot attention serves as an affinity measure between image features and slot vectors, effectively segmenting objects in the image under the assumption that the slots capture object representations. We enhance this framework by leveraging the cross-attention mask extracted from the adapter layers in the diffusion model as a self-supervisory signal to guide slot attention, and/or vice versa. These dual attention masks, one from slot attention and the other from the diffusion model, encode similar semantics regarding the relationship between slot representations and image features. Normally, there are as many adapter cross-attention masks as there are UNet blocks, but we focus only on the one from the third upsampling block (the second-to-last) since this layer is very close to the output and empirically provides an attention mask most aligned with the objects in the image. We denote the slot attention mask by $\mathbf{A}_{\text{SA}}$ and the diffusion attention mask with $\mathbf{A}_{\text{DM}}$. In both cases, the attention masks are computed through the dot product between queries and keys, and normalized over the query dimension. In slot attention, slots act as

queries and image features as keys, whereas in the attention mechanism of the diffusion model, this relationship is inverted. Formally we can write

$$\mathbf{A}_{\text{SA}} = \text{Softmax}\left(\frac{k_{\text{SA}}(\mathbf{f}_{\text{SA}})q_{\text{SA}}(\mathbf{S})^\top}{\sqrt{D}}\right) \qquad \mathbf{A}_{\text{DM}} = \text{Softmax}\left(\frac{k_{\text{DM}}(\mathbf{S})q_{\text{DM}}(\mathbf{f}_{\text{DM}})^\top}{\sqrt{D}}\right) \tag{4}$$

where q_{SA}, k_{SA}, q_{DM}, and k_{DM} are learnable linear functions, and $\mathbf{f}_{\text{SA}}$ and $\mathbf{f}_{\text{DM}}$ represent image features. The cross-attention layer in the adapter employs a multi-head structure, resulting in multiple attention masks. We average these masks over the head dimension and use the result as the diffusion attention mask $\mathbf{A}_{\text{DM}}$.

The slot attention mask, $\mathbf{A}_{\text{SA}} \in \mathcal{R}^{(h\cdot w)\times N}$, encodes how each pixel in the image features relates to the slot representations. Ideally, this attention mask should converge to an instance segmentation mask, with each slot representing a distinct object. Conversely, the diffusion attention mask, $\mathbf{A}_{\text{DM}} \in \mathcal{R}^{N\times(h\cdot w)}$, shows the inverse relationship. In the optimal scenario, we expect this attention map to converge to the transpose of the instance segmentation mask, enabling the diffusion model to generate the input image accurately. We formulate a guidance loss to enforce the alignment of these dual attention masks by

$$\mathcal{L}_{\text{guidance}} = \text{BCE}(\mathbf{A}_{\text{SA}}, \mathbf{A}_{\text{DM}}^\top), \tag{5}$$

where BCE is the binary cross-entropy loss. The overall training objective combines this guidance loss with the primary loss function, weighted by a hyperparameter λ:

$$\mathcal{L} = \mathcal{L}_{\boldsymbol{\theta}} + \lambda\mathcal{L}_{\text{guidance}} \tag{6}$$

There are different design choices for implementing the guidance loss: 1) guiding only the slot attention mask $\mathbf{A}_{\text{SA}}$ with $\mathbf{A}_{\text{DM}}$ (stopping gradient for diffusion model), 2) guiding only the diffusion attention mask $\mathbf{A}_{\text{DM}}$ with $\mathbf{A}_{\text{SA}}$ (stopping gradient for slot attention), and 3) joint guidance, that is, guiding both $\mathbf{A}_{\text{SA}}$ and $\mathbf{A}_{\text{DM}}$ simultaneously (no gradient stopping). So in the first two options, one attention mask serves as a pseudo ground-truth for the other, whereas in the last option, the two masks are jointly enforced for alignment.

Another alternative we have considered is guiding the attention masks through multiplication. In this scenario, there is no explicit auxiliary loss: each of the cross-attention masks from multiple attention heads in the adapter layer is simply multiplied by $\mathbf{A}_{\text{SA}}$ using the Hadamard product. Since each cross-attention mask ideally represents a part of the object bound by the corresponding slot, the multiplication confines the adapter attention to the region defined by the slot attention mask.

All the guidance alternatives described above aim, in one way or another, to enhance the alignment of the attention matrices which represent the same semantics in the architecture, thereby improving their alignment with the objects in the input image. We compare these different alternatives for attention mask guidance in the experiments section.

Table 1: **Architectural ablations on MOVi-E.** We examine the effects of architectural choices on segmentation and representation performance. We present block combinations on the left and register token choice on the right. Up, Down and Mid refer to all upsampling blocks, all downsampling blocks and middle block in the diffusion model.

	Conditioning Blocks					Register Token		
	Up+Down+Mid	Only Up	Only Down	Up+Mid	Up+Down	No Token	Slot Pooling	Feature Pooling
Segmentation (%)								
FG-ARI (↑)	56.89	57.38	**57.93**	57.39	56.45	54.38	56.27	**57.18**
mBO (↑)	39.59	43.05	40.20	39.96	**43.38**	40.07	41.65	**43.98**
mIoU (↑)	37.75	41.53	38.77	38.83	**41.86**	40.07	**40.10**	39.83
Representation								
Category (↑)	43.92	43.82	**45.88**	41.54	43.91	42.42	**43.54**	42.63
Position (↓)	1.92	1.82	**1.61**	1.92	1.72	1.89	**1.75**	1.78
3D B-Box (↓)	3.94	3.78	**3.48**	3.95	3.75	3.83	**3.77**	3.78

4 EXPERIMENTS

Datasets: Our evaluation framework covers both synthetic and real-world datasets, aligning with recent works in the field (Jiang et al., 2023; Wu et al., 2023b). We assess our method SlotAdapt on the synthetic MOVi-E dataset (Greff et al., 2022) and two widely-recognized real-world datasets:

VOC (Everingham et al., 2010) and COCO (Lin et al., 2014). While our primary focus is on leveraging pretrained diffusion models for real-world scenarios, we include MOVi-E in our evaluation due to its complexity, featuring scenes with up to 23 objects. This dataset serves as a challenging benchmark for object-centric learning in controlled environments. Both real-world datasets, VOC and COCO, have emerged as popular benchmarks for object discovery tasks. They present significant challenges due to their multi-object nature and the large number of foreground classes they contain—20 and 80, respectively. These datasets have been instrumental in recent evaluations of various object-centric learning methods on real-world images (Jiang et al., 2023; Wu et al., 2023b; Seitzer et al., 2023).

Implementation Details: Following the previous works (Jiang et al., 2023; Wu et al., 2023b), we use a convolutional backbone for MOVi-E and DINOv2 (Oquab et al., 2023) with ViT-B and a patch size of 14 as the encoder model for VOC and COCO. To serve as our decoder, we incorporate a pretrained Stable Diffusion (SD) model, v1.5 for MOVi-E; and COCO and v2.1 for VOC (Rombach et al., 2022), augmented with an additional cross-attention layer as adapter. For MOVi-E, we jointly optimize the image backbone, slot attention mechanism, and adapter layers. For VOC and COCO, we focus our training exclusively on the slot attention and adapter layers. We maintain consistency with previous work (Jiang et al., 2023; Wu et al., 2023b) in terms of dataset selection and preparation. Our models are trained for approximately 150K to 250K iterations. While this training duration is shorter compared to some previous works, we demonstrate that our approach achieves competitive performance.

Baselines: We compare SlotAdapt with unsupervised state-of-the-art object-centric methods. On the MOVi-E dataset, we compare with SLATE (Singh et al., 2021), SLATE$^+$, where SLATE's low capacity dVAE is replaced by a pre-trained VQGAN model (Esser et al., 2021) and Latent Slot Diffusion (LSD) (Jiang et al., 2023). On real-world datasets, we compare with Slot attention (SA) (Locatello et al., 2020), SLATE (Singh et al., 2021), DINOSAUR (Seitzer et al., 2023), LSD (Jiang et al., 2023) and SlotDiffusion (Wu et al., 2023b). In all our experiments on real-world datasets, we use a frozen DINO (Caron et al., 2021) as the visual encoder for all models.

Metrics: Following previous work (Locatello et al., 2020; Jiang et al., 2023), we employ a set of standard metrics to assess the performance of our method on unsupervised object segmentation. These include the foreground adjusted rand index (FG-ARI), mean intersection over union (mIoU), and mean best overlap (mBO). Our evaluation is performed on the slot attention masks, $\mathbf{A}_{\text{slot}}$, computed as in Equation 4. We use two different versions of the mIoU and mBO metrics: one computed over instance-level masks, and the other over class-level masks. Note that instance-level metrics account for whether objects of the same class in an image are differentiated as separate instances in the resulting segmentation; hence, they are more informative on the representational and generative capabilities of an object-centric learning method. On the other hand, the FG-ARI, a metric designed primarily for object discovery task, does not account for object masks larger than the ground-truth, which may be problematic, especially when assessing the generative capability.

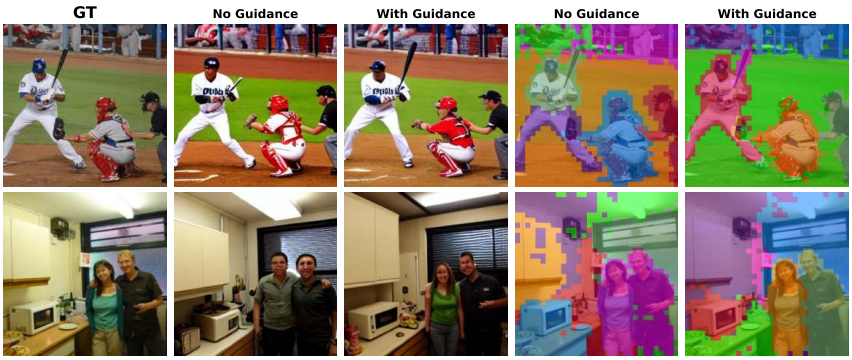

Figure 2: **Qualitative comparison: with vs. without guidance.** We visualize generated images and predicted segments on COCO dataset.

4.1 Ablation Studies

To assess the impact of our contributions, we have conducted a series of experiments. All the experiments in this ablation study are conducted with register token (slot pooling), joint guidance of

Table 2: **Evaluation of guidance strategies.** We present the segmentation performance on COCO for different guidance strategies. Joint guidance gives the best scores and significantly improves over no guidance option.

	FG-ARI	mBO^i	mBO^c	mIoU^i	mIoU^c
No guidance	42.3	31.5	34.8	31.7	38.5
Slot Guidance	41.2	33.4	36.9	33.1	37.9
DM Guidance	42.0	31.2	34.6	32.0	38.4
Joint Guidance	41.4	**35.1**	**39.2**	**36.1**	**41.4**
Multiplication Guidance	**43.3**	31.9	35.3	31.7	36.4

attention masks and conditioning through all downsampling and upsampling blocks, unless stated otherwise.

We first investigate the optimal block combination in the UNet architecture for conditioning the diffusion model on slots, where the options we consider are all downsampling blocks, all upsampling blocks, mid-block and their certain combinations. The results obtained on MOVi-E dataset are presented in the left part of Table 1. We observe that conditioning on either the upsampling blocks alone or both the downsampling and upsampling blocks yield superior performance, likely due to their proximity to the input and output in terms of structure and resolution.

Next, we evaluate the effect of incorporating a register token (Darcet et al., 2024) into the textual cross-attention layer in the diffusion model. The results obtained on MOVi-E are given in the right part of Table 1), where the options are no register token, slot pooling and feature pooling. We observe that inclusion of register token, with slot or feature pooling, yields consistent improvements in all metrics, including segmentation accuracy and performance in downstream tasks.

Lastly, we examine the effectiveness of our guidance strategies on COCO dataset in Table 2. We observe that joint guidance yields the best performance on all metrics except the FG-ARI score, where the improvements are substantial compared to the no-guidance case. For FG-ARI metric, multiplication guidance gives the best performance, which is a sufficient metric for object discovery but not necessarily for generative tasks. Moreover, in Fig. 2, we visually demonstrate the impact of the joint guidance strategy on the generated segmentation masks. We observe that the inclusion of guidance yields a significant improvement in segmentation quality, mitigating the part-whole hierarchy problem and producing segmentation masks better aligned with the objects in the scene, rather than with partial or fragmented objects. In turn, the generated (reconstructed) images are more faithful to the original input images, as also observed in Fig. 2.

4.2 Evaluation Results

All the evaluation experiments for SlotAdapt are conducted with register token, joint guidance and conditioning on all downsampling and upsampling blocks, unless stated otherwise.

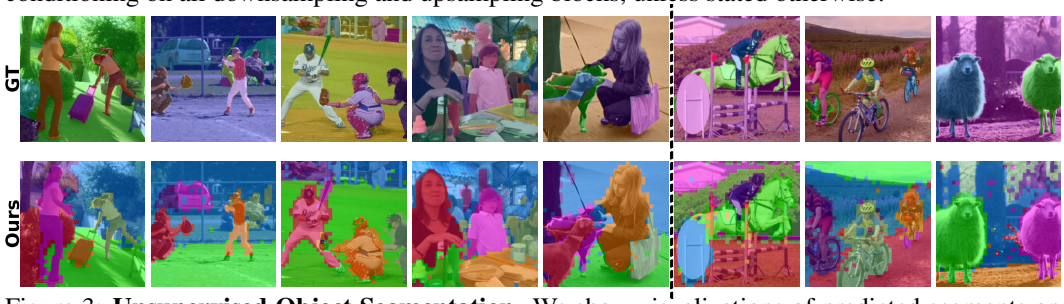

Figure 3: **Unsupervised Object Segmentation.** We show visualizations of predicted segments on COCO (left) and VOC (right). SlotAdapt successfully binds distinct instances belonging to the same class.

Synthetic Dataset: We evaluate the object discovery and segmentation performance of our method on MOVi-E in comparison to the baseline methods in Table 3, based on the FG-ARI, instance-level mBO and instance-level mIoU metrics. To measure the representational quality of the learned slots for downstream tasks, we implement a straightforward approach using a 2-layer MLP to predict discrete categories, object positions and 3D bounding boxes. We also display the resulting prediction

Table 3: **Comparative evaluation on MOVi-E:** (Left) Segmentation results, (Right) Representation assessment: We evaluate slots through predictive probing. Spatial attributes (position, 3D bounding box) are assessed via MSE (mean squared error), while categorical predictions are assessed by classification accuracy.

Segmentation	SLATE	SLATE$^+$	LSD	Ours
mBO (↑)	30.17	22.17	38.96	**43.38**
mIoU (↑)	28.59	20.63	37.64	**41.85**
FG-ARI (↑)	46.06	45.25	52.17	**56.45**

Representation	SLATE	SLATE$^+$	LSD	Ours
Position (↓)	2.09	2.15	1.85	**1.77**
3D B-Box (↓)	3.36	3.37	**2.94**	3.75
Category (↑)	38.93	38.00	42.96	**43.92**

accuracies in Table 3. We observe significant improvements in almost all metrics, particularly with an approximate 10% enhancement in object discovery and segmentation accuracy.

Real-World Dataset: We evaluate the object discovery and segmentation performance of our method on the COCO and VOC datasets in Table 4, using the FG-ARI, instance-level mBO and class-level mBO metrics, in comparison to baseline methods including StableLSD (Jiang et al., 2023), SlotDiffusion (Wu et al., 2023b), and DINOSAUR (Seitzer et al., 2023).

We observe that SlotAdapt, when used with guidance, outperforms all the baselines for almost all metrics, including the highly competitive DINOSAUR method. The only exception is the class-level mBO score on VOC, where SlotAdapt performs worse than SlotDiffusion. Notably, the improvement in mBO scores on COCO is particularly significant, about 10% better than the next best baseline. The impact of guidance is also substantial on COCO, which is a more challenging and much larger dataset with complex multi-object scenes and varying object sizes, when compared to VOC.

In Fig.4, we visually compare the segmentation results of LSD, SlotDiffusion and SlotAdapt on COCO. We observe that SlotAdapt successfully differentiates individual object instances as evidenced by its superior instance-level mBO score, whereas LSD and Slot Diffusion struggle with this challenge. Moreover, SlotAdapt produces more complete segmentations of objects without dividing them into parts, which is reflected by its higher FG-ARI score. These results highlight the robustness and versatility of SlotAdapt in handling the complexities of real-world data.

Generation and Compositional Editing: We first demonstrate the ability of our model to generate realistic images on COCO in Fig. 5. We see that, when conditioned on slots, our model can reconstruct the input image with high quality, realism and notable fidelity. Regarding compositional generation and editing capabilities, Figure 6 shows a series of image edits by modifying input slots, including object replacement, removal and addition. We observe that the editing operations are highly successful and seamless with only slight yet realistic changes to the image background, while all maintaining high quality.

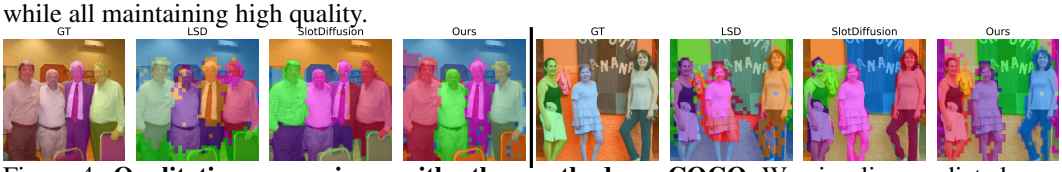

Figure 4: **Qualitative comparisons with other methods on COCO.** We visualize predicted segments of SlotAdapt in comparison to LSD and SlotDiffusion. SlotAdapt can more effectively differentiate between object instances of the same class compared to other methods.

Figure 5: **Generation Results.** We show sample images reconstructed by SlotAdapt conditioned on slots on COCO (left) and VOC (right). SlotAdapt generates reconstructions highly faithful to the original input images.

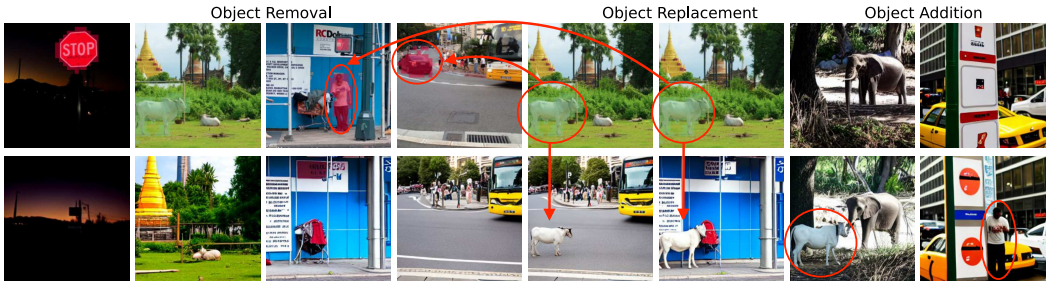

Figure 6: **Compositional Editing.** We demonstrate object removal, replacement and addition edits on COCO images by using slots. Removing highlighted slots (top row) yields realistic and successful generations (first 4 examples). Replacing highlighted objects in the 3rd and 4th images with the cow object from the 5th and 6th images results in highly accurate edits, yet with small changes in the original images. Finally, adding the cow (5th image) and the person (3rd image) slots to the last two images, respectively, generates meaningful examples of complex scenes.

Table 4: **Unsupervised object segmentation on real-world datasets.** We compare SlotAdapt with state-of-the-art methods on VOC (left) and COCO (right). We present two versions of our method, with and without guidance loss.

VOC	FG-ARI	mBO^i	mBO^c
SA + DINO ViT	12.3	24.6	24.9
SLATE + DINO ViT	15.6	35.9	41.5
DINOSAUR	23.2	43.6	50.8
LSD	18.7	40.5	43.5
SlotDiffusion	17.8	50.4	**55.3**
Ours	28.8	**51.6**	52.0
Ours + Guidance	**29.6**	51.5	51.9

COCO	FG-ARI	mBO^i	mBO^c
SA + DINO ViT	21.4	17.2	19.2
SLATE + DINO ViT	32.5	29.1	33.6
DINOSAUR	34.3	32.3	38.8
LSD	33.8	27.0	30.5
SlotDiffusion	37.2	31.0	35.0
Ours	**42.3**	31.5	34.8
Ours + Guidance	41.4	**35.1**	**39.2**

5 Conclusion

We have targeted the object-centric learning problem, particularly on complex real-world images. To this end, we have presented a method that combines slot attention with pretrained diffusion models. Our architecture, SlotAdapt, has three main novelties; adapters in the diffusion model for slot conditioning, the use of a register token to represent background in images, and attention guidance to align slot attention with cross attention from the diffusion model. We have conducted extensive experiments, particularly on COCO dataset, to validate our approach. Our experiments show that leveraging the generative capabilities of pretrained diffusion models is crucial for tackling the challenging tasks such as object discovery, unsupervised segmentation and compositional generation and editing on complex real-world images. Our method takes a step forward in achieving this goal, outperforming the state of the art methods on the challenging COCO dataset without relying on any external supervision in object discovery, segmentation and, particularly compositional generation and editing. We are the first to present successful compositional editing experiments on the COCO dataset.

A limitation of our method from the perspective of compositional generation is that the edited or reconstructed images, when conditioned on slots, may exhibit slight changes with respect to the source image, though the generations are mostly highly realistic and of very good quality. One remedy for this can be the incorporation of additional training objectives to enforce fidelity to the input. Another related issue is how to make use of this framework for image editing in practice, which needs further adjustments to the architecture possibly to accept user prompts and associate them with slot representations. Lastly, our method still has issues for under- and over-segmentation, which may potentially be mitigated through slot merging and splitting. Additionally, our approach relies on pre-trained diffusion models that are predominantly trained on real-world data, which may limit their adaptability to synthetic domains. While the model performs well on natural images, its effectiveness may be reduced when handling synthetic datasets such as CLEVR and CLEVRTex Johnson et al. (2017).

REFERENCES

Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. Systematic generalization: What is required and can it be learned? In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2019.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Léon Bottou. From machine learning to machine reasoning. *Machine Learning*, 94(2):133–149, 2014.

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 *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8780–8794, 2021.

Gamaleldin Fathy Elsayed, Aravindh Mahendran, Sjoerd van Steenkiste, Klaus Greff, Michael Curtis Mozer, and Thomas Kipf. SAVi++: Towards end-to-end object-centric learning from real-world videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

SM Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Geoffrey E Hinton, et al. Attend, infer, repeat: Fast scene understanding with generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 12873–12883, 2021.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision (IJCV)*, 2010.

Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

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 *Proc. of the International Conf. on Machine Learning (ICML)*, 2019.

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

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3749–3761, 2022.

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.

Geoffrey Hinton. Some demonstrations of the effects of structural descriptions in mental imagery. *Cognitive Science*, pp. 231–250, 1979.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshops*, 2021.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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.

Baoxiong Jia, Yu Liu, and Siyuan Huang. Improving object-centric learning with query optimization. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2023.

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

Jindong Jiang, Sepehr Janghorbani, Gerard De Melo, and Sungjin Ahn. SCALOR: Generative world models with scalable object representations. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2019.

Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 8563–8601, 2023.

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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Whie Jung, Jaehoon Yoo, Sungjin Ahn, and Seunghoon Hong. Learning to compose: Improving object centric learning by injecting compositionality. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2024.

Daniel Kahneman, Anne Treisman, and Brian J. Gibbs. The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, pp. 175–219, 1992.

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 *Proc. of the International Conf. on Learning Representations (ICLR)*, 2021.

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

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, 2017.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pp. 740–755. Springer, 2014.

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 *Proc. of the International Conf. on Learning Representations (ICLR)*, 2019.

Zhixuan Lin, Yi-Fu Wu, Skand Peri, Bofeng Fu, Jindong Jiang, and Sungjin Ahn. Improving generative imagination in object-centric world models. In *Proc. of the International Conf. on Machine Learning (ICML)*, 2020.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 11525–11538, 2020.

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. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2021.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proc. of the Conf. on Artificial Intelligence (AAAI)*, volume 38, pp. 4296–4304, 2024.

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 *Proc. of the International Conf. on Machine Learning (ICML)*, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. In *Transactions on Machine Learning Research (TMLR)*, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, pp. 1–14, 2022b.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 109(5):612–634, 2021.

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 *Proc. of the International Conf. on Learning Representations (ICLR)*, 2023.

Gautam Singh, Fei Deng, and Sungjin Ahn. Illiterate dall-e learns to compose. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2021.

Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Krishnakant Singh, Simone Schaub-Meyer, and Stefan Roth. Guided latent slot diffusion for object-centric learning. *arXiv preprint arXiv:2407.17929*, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. of the International Conf. on Machine Learning (ICML)*, 2015.

Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental Science*, 2007.

Yanbo Wang, Letao Liu, and Justin Dauwels. Slot-vae: Object-centric scene generation with slot attention. In *International Conference on Machine Learning*, pp. 36020–36035. PMLR, 2023.

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. Structured world modeling via semantic vector quantization. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2024.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. SlotFormer: Unsupervised visual dynamics simulation with object-centric models. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2023a.

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

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

A APPENDIX

In this appendix, we provide more details and results about our work. It is organized into three parts:

- Training and Architectural Details (section A.1): We explain how we trained our model SlotAdapt and describe its structure in depth.
- More Examples (section A.2): We show additional visual examples of SlotAdapt's performance in different settings.
- Comparison with GLASS (section A.3): We compare SlotAdapt with GLASS, a similar concurrent study, to show how our work fits into current research.
- Additional Experiments (section A.4) We conduct additional experiments by following reviewers' suggestions
- Additional Evaluations (section A.5) We conduct additional evaluations by following reviewers' suggestions
- New visual examples (section A.6) We provide new visual examples with CFG value of 1.3 and different seed values

A.1 TRAINING AND ARCHITECTURAL DETAILS

This section provides comprehensive information on the architectural components and training procedures of our model. Moreover, we include the code and training scripts in the supplementary material as a reference.

A.1.1 ARCHITECTURAL DETAILS

Figure 1 in the main text illustrates our training pipeline. Below, we detail the key components:

Diffusion Model: We initialize a UNet denoiser and VAE from pretrained diffusion models (Rombach et al., 2022). Specifically:

- MOVi-E (Greff et al., 2022) and COCO Lin et al. (2014) datasets: Stable Diffusion v1.5
- VOC (Everingham et al., 2010): Stable Diffusion v2.1

After initialization, we inject adapter layers following each downsampling and upsampling block in the UNet. Each adapter layer comprises a cross-attention mechanism and a feedforward network, both preceded by layer normalization.

VAE: We employ pretrained VAEs, maintaining consistency with the diffusion model versions: Stable Diffusion v1.5 VAE for MOVi-E and COCO; Stable Diffusion v2.1 VAE for VOC

Visual Encoder Backbone:

- MOVi-E: We utilize a custom CNN backbone encoder. It consists of 4 downsampling blocks, 1 middle block, and 4 upsampling blocks, with channel multipliers [1,1,2,4] and a base channel count of 128. Each block contains 2 residual layers and the overall network's output channels 128 channels.
- COCO and VOC: We employ a frozen DINOv2 (Oquab et al., 2023) model with a ViT-B backbone (patch size 14).

The extracted feature maps maintain a consistent resolution of 32×32 across all datasets.

Slot Attention: Our implementation varies by dataset:

- MOVi-E: We adopt the LSD (Jiang et al., 2023) configuration (slot size: 192, iterations: 3, slots: 24). We append four linear projectors to align slot dimensions with adapter attention layers.

- COCO and VOC: We use a slot size of 768 and 7 slots, with similar output projector layers as in MOVi-E.

For all datasets, we use a linear layer to project either pooled visual backbone features or averaged slot vectors to a 768-dimensional space to match the text cross-attention dimensions in the diffusion model. We use slot averaging for MOVi-E and COCO datasets, and feature pooling for the VOC dataset, as they perform slightly better.

Training Details:

- Hardware: 2 NVIDIA A40 GPUs
- Batch sizes: 32 (MOVi-E), 30 (VOC), 32 (COCO)
- Training iterations: 200K (MOVi-E), 250K (COCO), 190K (VOC)
- Optimization: AdamW optimizer, constant learning rate, FP16 precision
- λ: In first 40K iterations, we use 0, from 40K-50K, it is increased from 0 to the 0.025, then, we use 0.025. Since the slots and adapter layers are initialized completely random, we opt to wait for 40K iterations so that the attention masks have meaningful connections.

We will release our code upon publication and have included it in the supplementary materials for reference.

Experimental Details: For our experiments, we first select the arhictectural details (conditioning place and register token type) on MOVi-E dataset. Then, we migrate that architecture to the real-world setup.

Dataset Details: For **MOVi-E**, we follow the train-test split in Singh et al. (2022) and use 256×256 resolution for both diffusion model and slot attention. For **COCO**, we follow Seitzer et al. (2023) to train on the same training set with DINOSAUR, LSD and Slot Diffusion, which consists of 118,287 images for training and 5000 for validation. For **VOC**, we follow Seitzer et al. (2023) to train on the "trainaug" set so that the training sets are the same as DINOSAUR, LSD and Slot Diffusion. The training set contains 10,582 images and validation set contains 1449 images. For COCO and VOC, we use random horizontal flip in training as data augmentation and we use 256×256 for diffusion model and 448×448 for slot attention.

A.2 Additional Qualitative Examples

In this section, we present additional qualitative examples to further illustrate the capabilities of our proposed SlotAdapt model. These results complement the main paper by providing a more comprehensive view of our model's performance across various tasks and datasets.

Figure 7: **Unsupervised Object Segmentation.** We show visualizations of predicted segments on COCO.

Figure 8: **Unsupervised Object Segmentation.** We show visualizations of predicted segments on VOC.

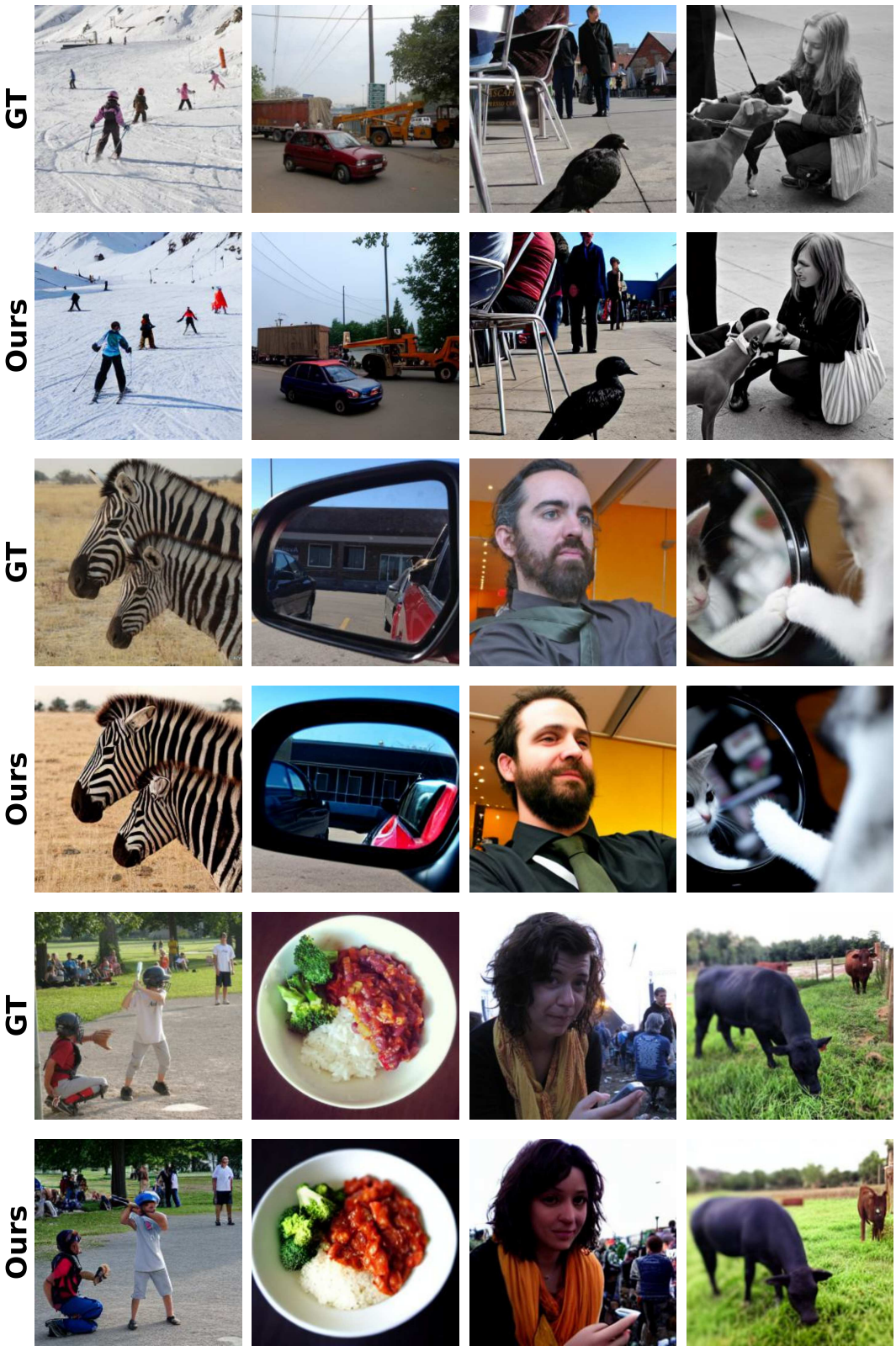

Figure 9: **Generation Results.** We show sample images reconstructed by SlotAdapt conditioned on slots on COCO.

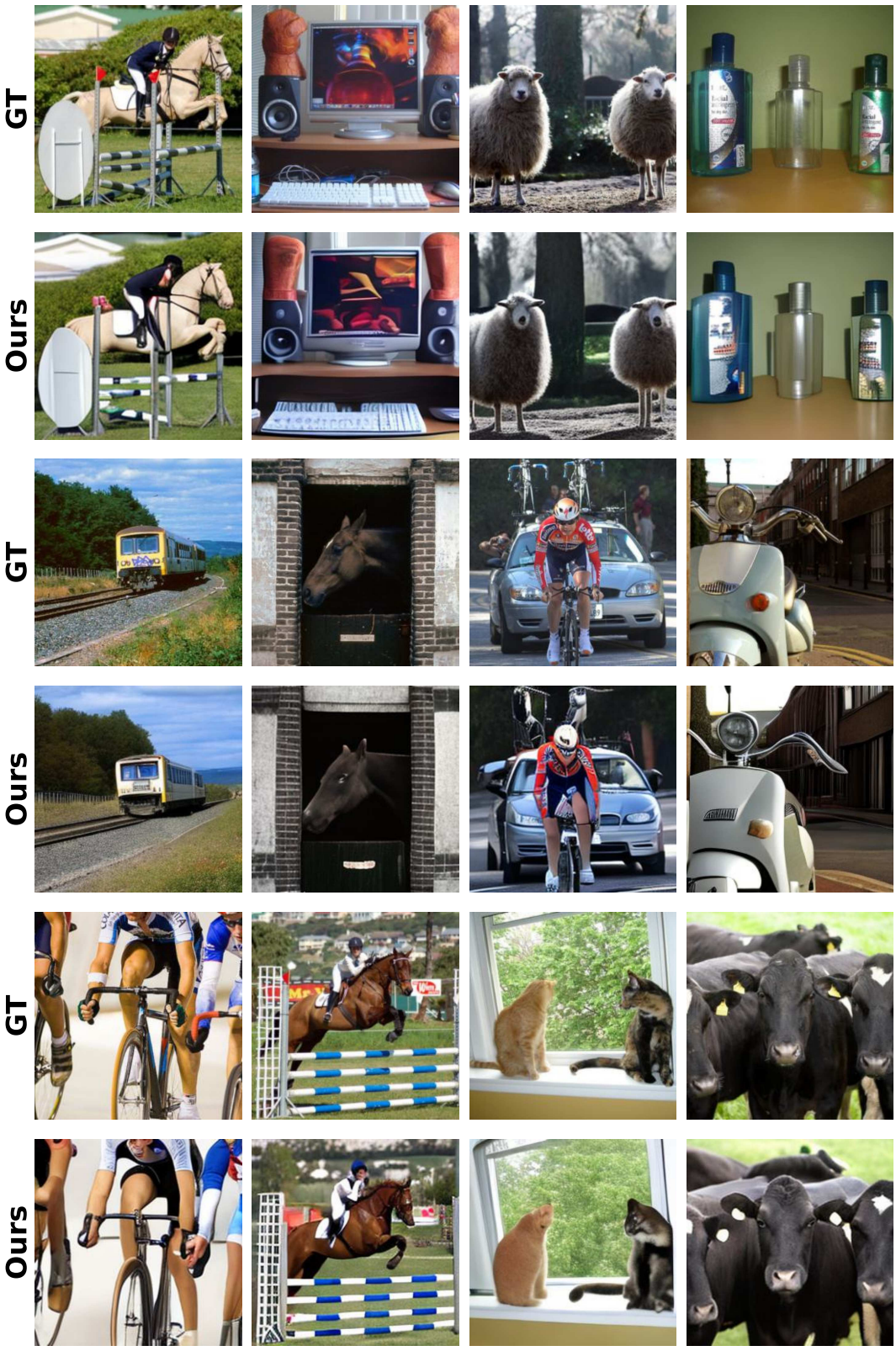

Figure 10: **Generation Results.** We show sample images reconstructed by SlotAdapt conditioned on slots on VOC.

Figure 11: **Segmentation comparisons with other methods** We show visualizations of predicted segments on COCO dataset. Compared to other models, our model predicts more complete masks without segmenting objects into parts.

Figure 12: **Generation comparisons with other methods.** We show visualizations of generated images on COCO dataset. Compared to other models, our model can generate better reconstructions.

Figure 13: **Segmentation comparisons with other methods.** We show visualizations of predicted segments on VOC dataset. Compared to other models, our model predicts more complete masks without segmenting objects into parts.

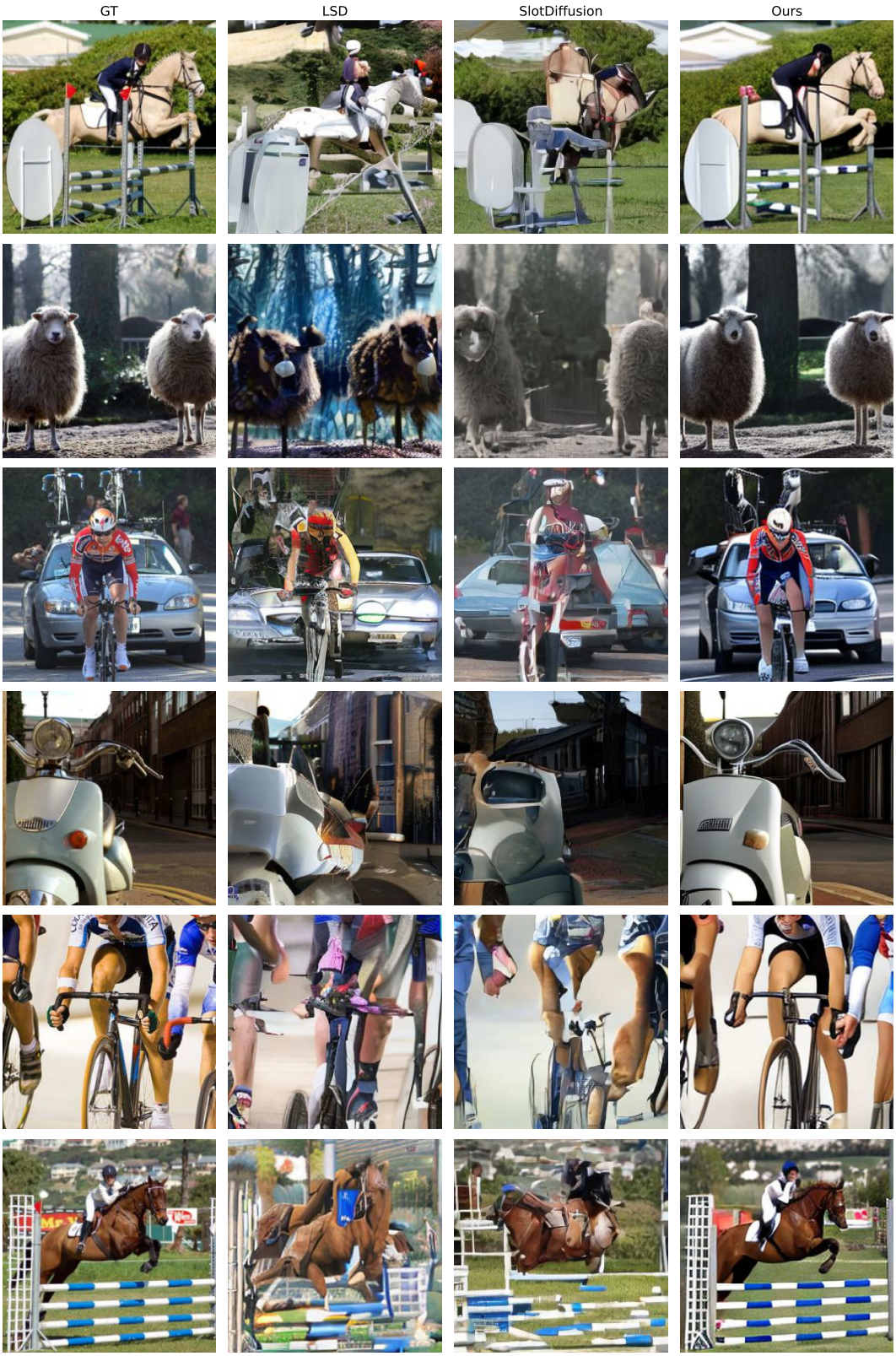

Figure 14: **Generation comparisons with other methods.** We show visualizations of generated images on VOC dataset. Compared to other models, our model can generate better reconstructions.

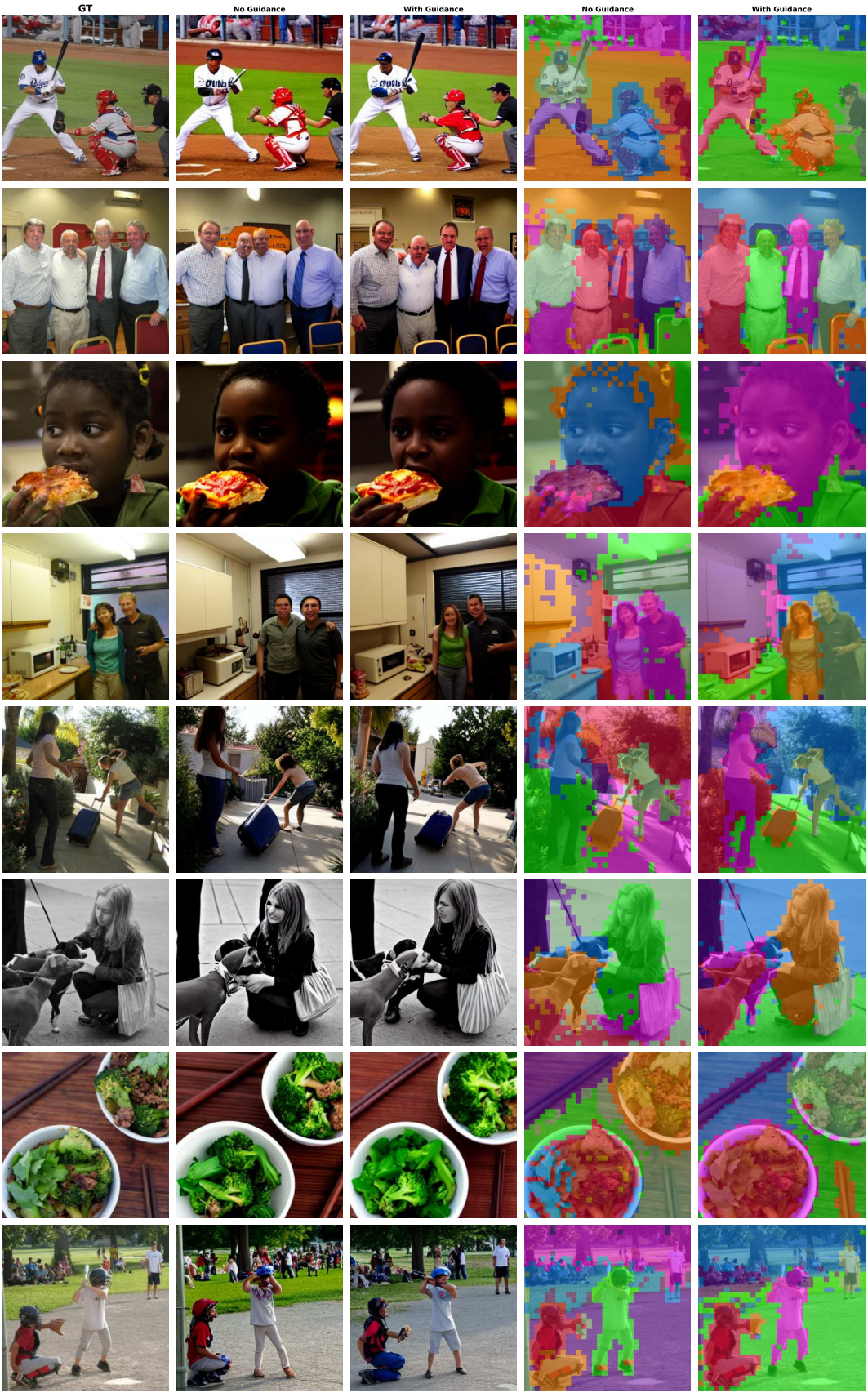

Figure 15: **Qualitative comparison: with vs. without guidance.** We visualize generated images and predicted segments on COCO dataset. Guidance helps to generate better aligned objects and to solve the "part-whole" hierarchy problem in object segmentation task.

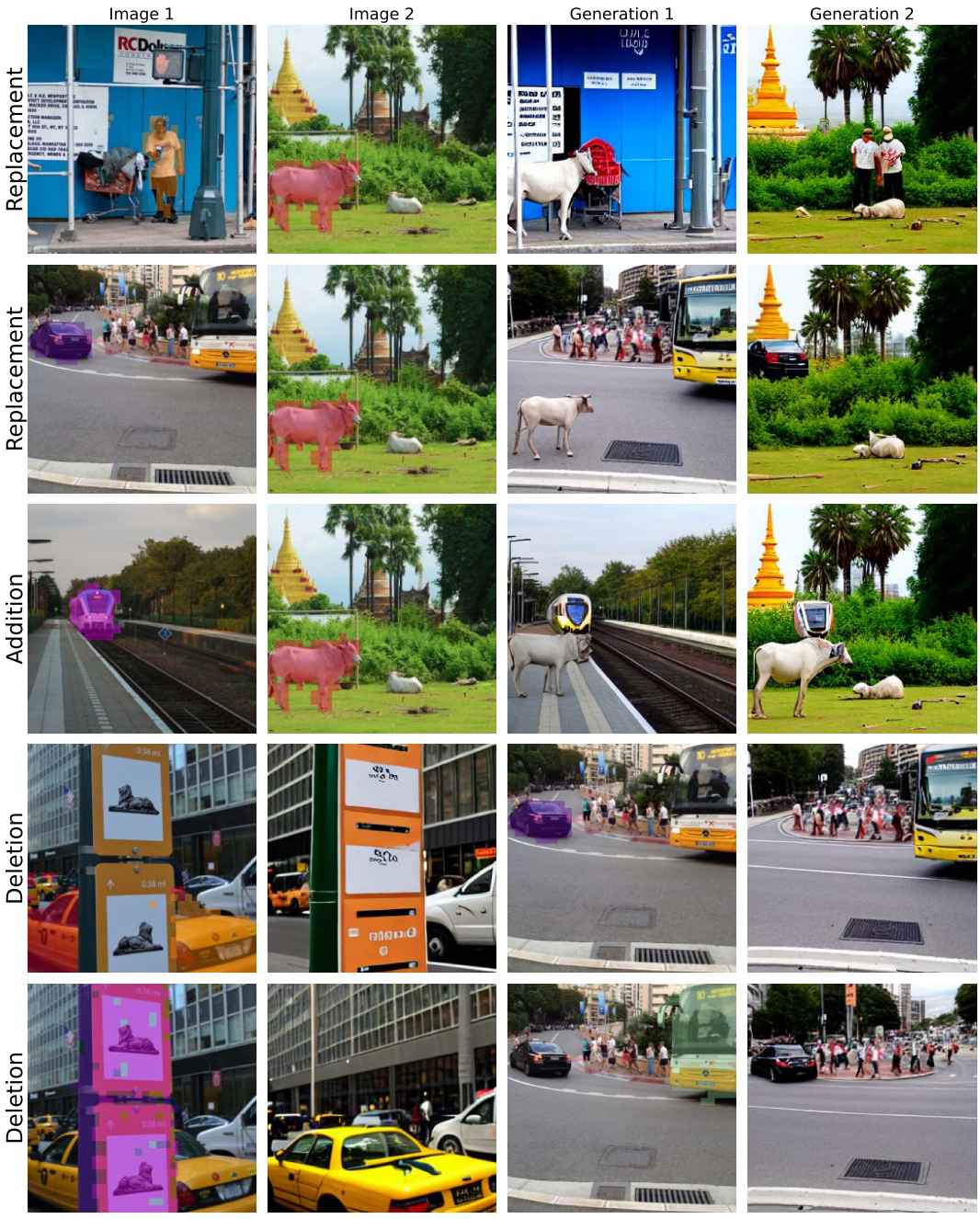

Figure 16: **Compositional Generations and Editing.** We show visualizations of generated images when the slots are manipulated. In the first two rows, we exchange the highlighted slots. In the middle row, we simply add the highlighted slot to the other image. In the last two rows, we delete the highlighted slots to manipulate the image. For all scenarios, our model can successfully generate realistic images in complex scenarios. This shows our model's ability to compositional generation and editing.

A.3 COMPARISON WITH GLASS

This section presents a comprehensive comparison between our method and the concurrent work GLASS (Singh et al., 2024), highlighting both quantitative and qualitative differences.

Methodological Distinctions: GLASS leverages extra information, such as class labels or image captions, to enhance its performance. While this approach yields certain advantages, it introduces limitations. Primarily, GLASS struggles to differentiate between instances of the same class due to its reliance on semantic masks as pseudo ground truth. Additionally, the need for additional information restricts the applicability of GLASS in fully unsupervised scenarios. In contrast, our method operates without any external supervision, successfully segments individual instances even within the same class, and captures nuanced scene information without relying on pre-defined semantic categories.

Figure 17 provides visual comparisons on both COCO and VOC datasets. Our model demonstrates superior performance in instance differentiation, accurately segmenting multiple instances of the same class (as shown in rows 1, 2, and 4). Furthermore, it excels in scene understanding, capturing meaningful elements such as trees and house roofs (row 3) without explicit labeling.

Table 5 presents a quantitative comparison of our method with GLASS. In terms of semantic overlap, GLASS excels, which is attributable to its use of semantic masks for supervision. However, on complex datasets like COCO, which features multiple instances per image, SlotAdapt achieves comparable performance without any additional supervision. These results underscore the robustness and versatility of our unsupervised approach, particularly in handling complex, multi-instance scenarios. Our method demonstrates that high-quality object segmentation can be achieved without relying on external supervision, offering a more flexible and generalizable solution for diverse image understanding tasks.

Table 5: Unsupervised object segmentation comparisons with the concurrent work GLASS on VOC (left) and COCO (right). We would like to point out that both GLASS and GLASS† use extra supervision such as class labels or image caption.

PASCAL VOC	FG-ARI	mBO^i	mBO^c
Ours	28.8	51.6	52.0
Ours + Guidance	**29.6**	51.5	51.9
GLASS†	—	**60.4**	**68.4**
GLASS	—	58.1	36.1

MS COCO	FG-ARI	mBO^i	mBO^c
Ours	**42.3**	31.5	34.8
Ours + Guidance	41.4	35.1	39.2
GLASS†	—	34.3	45.2
GLASS	—	**35.3**	**46.3**

Figure 17: **Unsupervised Object Segmentation.** We show visualizations of predicted segments for SlotAdapt vs. GLASS on real world datasets (VOC and COCO).

Figure 18: **Generation results.** We show visualizations of generated images by SlotAdapt vs. GLASS on COCO and VOC.

A.4 Additional Experiments

In this section, we have included the new experiments that are asked by the reviewers during the rebuttal period.

Better Segmentation Model: In this part, we replace Slot Attention with BOQ-SA, which is an improved version of slot attention where the slot initialization is also optimized, Jia et al. (2023). The results show that using BOQ-SA instead of Slot Attention results in modest improvements across several metrics:

Slot Number: As asked by the reviewers, the slot number is selected by following the previous works Seitzer et al. (2023); Wu et al. (2023b); Jiang et al. (2023), which is fixed to the maximum number of objects in the scene in a specific dataset. We present the results with 80 slots versus 7 slots (original). Increasing the slot number to such a high number results in poor results.

Starting Guidance Loss from the start: We also experiment with starting of guidance loss, where the original idea was starting the guidance loss after some iterations so that the slots and the adapters have a knowledge of the objects. Below, we present results where the guidance loss is started in the beginning without waiting slots and adapters learn meaningful features. The results show that starting guidance loss after some iterations leads to better performance.

Additional Token for Slot Attention as a register token: Following the reviewer's suggestion, we investigated an alternative approach to capturing global scene information in our architecture. Instead of using the average of all slot tokens as a global register token, we modified the slot attention module to include an additional dedicated slot specifically designed to capture global scene context. This dedicated slot was then directly used in the cross-attention layer. Our experimental results revealed interesting trade-offs: while the dedicated global slot approach improved foreground segmentation quality, it showed slightly decreased performance in semantic segmentation and generation quality. We hypothesize that this performance difference stems from the competitive nature of attention in the slot attention module—the dedicated global slot must compete with object slots for attention weights, potentially limiting its capacity to capture comprehensive scene information compared to our original approach of averaging all slot representations. This suggests that using the collective information from all slots through averaging provides a more robust global representation than a single learned global token.

A.5 Additional Evaluations

In this section, we provide further evaluations in terms of image quality metrics such as FID Heusel et al. (2017) and KID Bińkowski et al. (2018). We evaluate the reconstructions by our method and the baselines methods and the compositional generation ability of each method. Further, we conduct an ablation study for classifier-free guidance value, which results into improvement results.

Reconstruction Ability: We evaluate our method and the baseline methods, SlotDiffusion Wu et al. (2023b) and LSD Jiang et al. (2023) using FID Heusel et al. (2017) and KID Bińkowski et al. (2018). The results show that our methods generation ability is much better and leads to more realistic images.

Compositional Generation Ability: We evaluate the ability of compositional generation of our method and the baseline methods using the method used by SlotDiffusion Wu et al. (2023b). SlotDiffusion proposes an idea where the slots in the batch are randomly mixed which results randomly composed slots from each concept's library to generate novel samples. We follow the original evaluation code shared by the SlotDiffusion authors. The results show that our method's compositional generation ability is superior compared to both SlotDiffusion Wu et al. (2023b) and LSD Jiang et al. (2023).

Classifier-free Guidance: We conduct an ablation study for classifier-free guidance value, which simply increases the effect of the conditioning in diffusion models Ho & Salimans (2022). We

Table 6: Comparison of segmentation methods on real-world data.

Method	FG-ARI	mBO^i	$mIoU^i$	mBO^c	$mIoU^c$
Slot Attention	42.3	31.5	31.7	34.8	38.5
BOQ-SA	42.2	31.2	32.551	35.266	37.82

Table 7: Comparison of different slot numbers.

Method	FG-ARI	Instance		Semantic		FID	KID
		mBO	mIoU	mBO	mIoU		
80 slots	18.1	23.2	26.4	28.6	32.9	114.236	64.610
7 slots	41.4	35.1	36.1	39.2	41.4	10.857	0.388

find that CFG value of 1.3 results in the best results in terms both FID Heusel et al. (2017) and KID Bińkowski et al. (2018).

Table 8: Comparison of different training starting points.

Method	FG-ARI	Instance		Semantic	
		mBO	mIoU	mBO	mIoU
Start from 0	37.85	32.65	33.99	36.059	39.254
Original (start after 40K)	41.4	35.1	36.1	39.2	41.4

Table 9: Comparison of different global information capturing strategies.

Method	FG-ARI	Instance		Semantic		FID	KID
		mBO	mIoU	mBO	mIoU		
Additional Slot Token	43.8	31.9	32.4	35.5	37.3	11.212	0.431
Slot Average Token	42.3	31.5	34.8	34.8	38.5	10.857	0.388

Table 10: **Reconstruction Ability** Comparison of FID and KID scores across different methods.

Method	FID	KID×1000
Ours	10.857	0.388
LSD	35.537	19.086
SlotDiff	19.448	5.852

Table 11: **Compositional Generation Ability** Comparison of FID and KID scores across different methods.

Method	FID	KID×1000
Ours w/ guidance	40.568	34.381
LSD	167.232	103.482
SlotDiff	64.213	57.309

Table 12: Effect of CFG values on generation quality metrics.

CFG Value	FID	KID×1000
7.5	21.350	6.271
5.0	17.558	4.236
2.5	13.734	2.151
2.0	12.427	1.424
1.5	11.041	0.590
1.4	10.880	0.459
1.3	**10.857**	**0.388**
1.2	11.057	0.492

A.6 NEW VISUALIZATIONS

In this section, we provide new qualitative examples where we used new classifier-free guidance (CFG) value of 1.3, as it is shown to perform better. Also, we provide visuals where we change the CFG value and the seed value. We will replace the generated visuals with those obtained by a CFG of 1.3 in the final version of the main paper, where CFG value of 2.5 was used, if accepted.

Figure 19: This is the corresponding figure for Figure 2

Figure 20: This is the corresponding figure for Figure 5

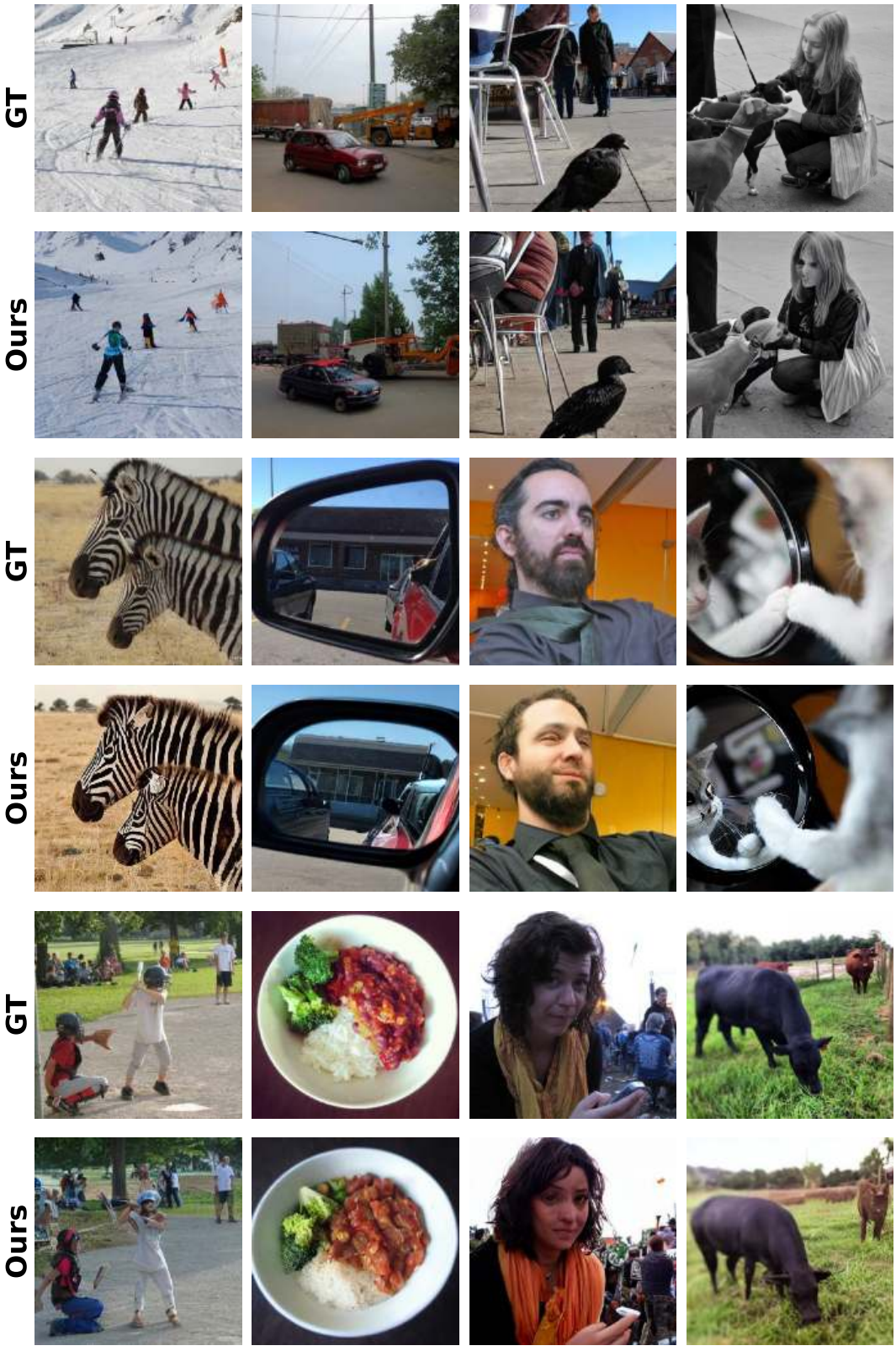

Figure 21: This is the corresponding figure for Figure 9

Figure 22: This is the corresponding figure for Figure 12

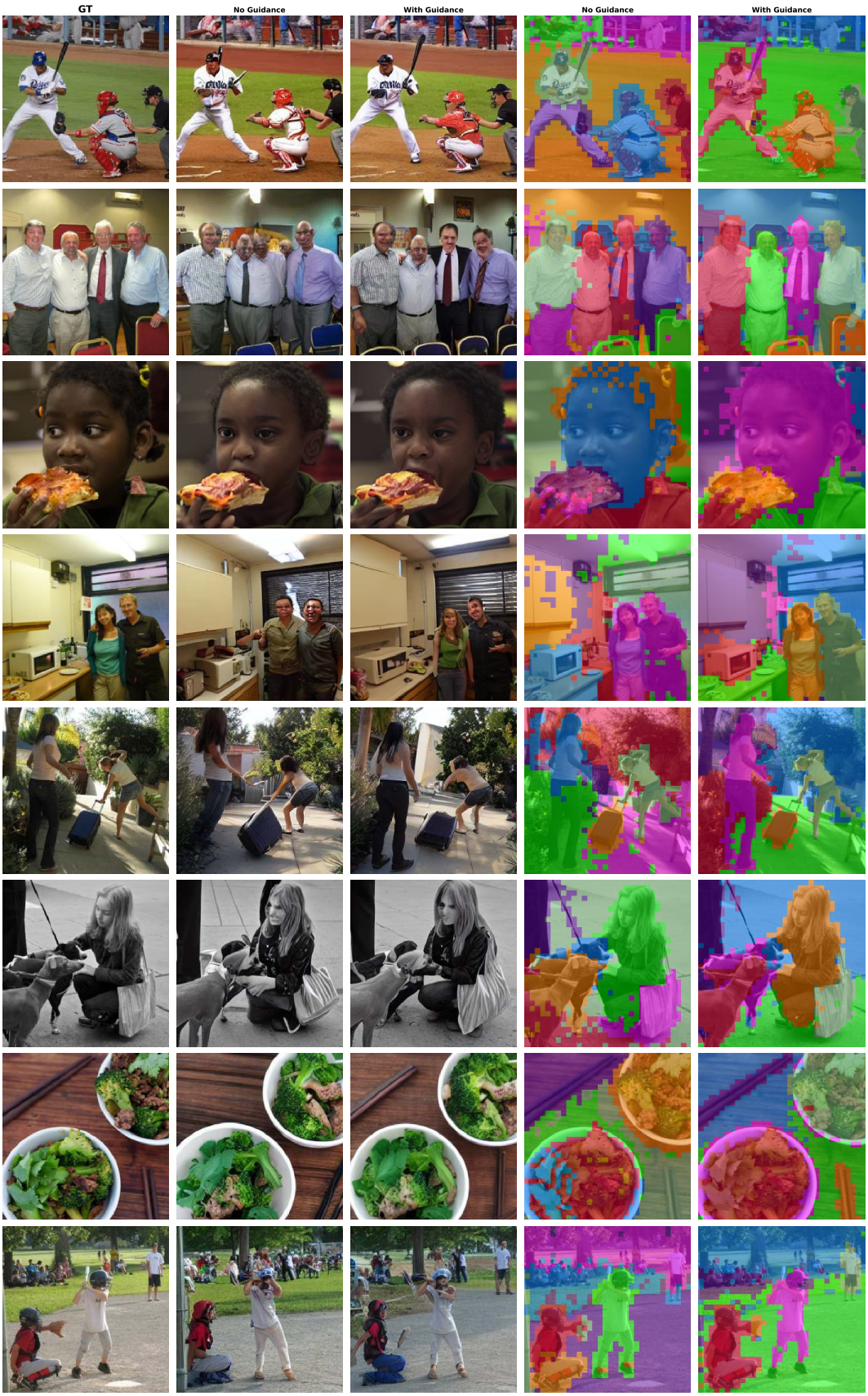

Figure 23: This is the corresponding figure for Figure 15

Figure 24: This is the corresponding figure for Figure 18

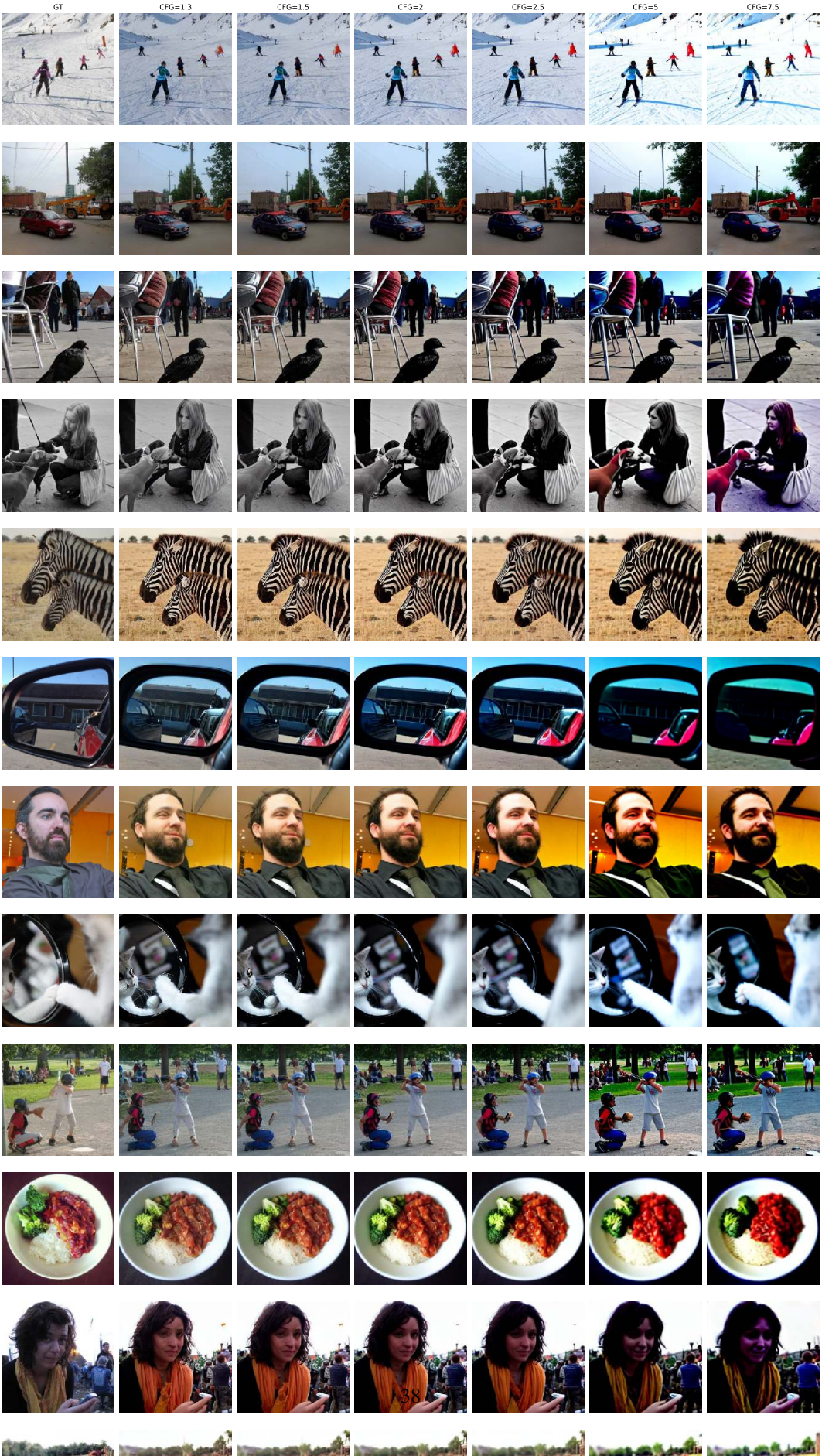
GT
CFG=1.3
CFG=1.5
CFG=2
CFG=2.5
CFG=5
CFG=7.5

Figure 26: **Seed Value** As asked by the reviewers, we examine how seed value effects the slot attention segmentation process.