CVPR #12908

GLASS: Guided Latent Slot Diffusion for Object-Centric Learning

Anonymous CVPR submission

Paper ID 12908

Abstract

Object-centric learning aims to decompose an input image into a set of meaningful object files (slots). These latent object representations enable a variety of downstream tasks. Yet, object-centric learning struggles on real-world datasets, which contain multiple objects of complex textures and shapes in natural everyday scenes. To address this, we introduce ***G****uided* ***La****tent* ***S****lot Diffu****s****ion (GLASS), a novel slot attention model that learns in the space of generated images and uses semantic and instance guidance modules to learn better slot embeddings for various downstream tasks. Our experiments show that GLASS surpasses state-of-the-art slot attention methods by a wide margin on tasks such as (zero-shot) object discovery and conditional image generation for real-world scenes. Moreover, GLASS enables the first application of slot attention to compositional generation of complex, realistic scenes.*[*]

1. Introduction

Humans perceive a scene as a collection of objects [35]. Such a decomposition of the scene into objects makes humans capable of higher cognitive tasks like control, reasoning, and the ability to generalize to unseen experiences [25]. Building on these ideas, object-centric learning (OCL) aims to decompose a scene into compositional and modular symbolic components. OCL methods bind these components to latent (neural) representations, enabling such models to be applied to tasks like causal inference [59], reasoning [2], control [4], and out-of-distribution generalization [15].

Slot attention models [45], a popular class of OCL methods, decompose an image into a set of latent representations where each element, called slot, competes to represent a certain part of the image. Slot attention methods can be categorized as form of representation learning, where the representation (slots) facilitates various downstream tasks such as property prediction [15], image reconstruction [32], image editing [71], and object discovery [60]. However, numerous promising slot attention methods [45, 62, 63] have remained limited to synthetic and simple datasets [16, 26, 33, 37]. Some recent methods [32, 36, 60, 71] use powerful modern encoder [7, 51] and decoder networks [56] to scale to complex real-world imagery [21, 44]. Yet, these models remain restricted to object discovery, lacking the versatility to reconstruct or perform compositional generation of realistic images. Moreover, the quality of the obtained slot representations remains limited as witnessed by both qualitative and quantitative results, which show the slots to suffer from the issue of over-segmentation (segmenting an object into multiple slots), under-segmentation (segmenting multiple objects into one slot), or imprecise object boundaries. This over- and under-segmentation issue is also known as the part-whole hierarchy ambiguity [29–31].

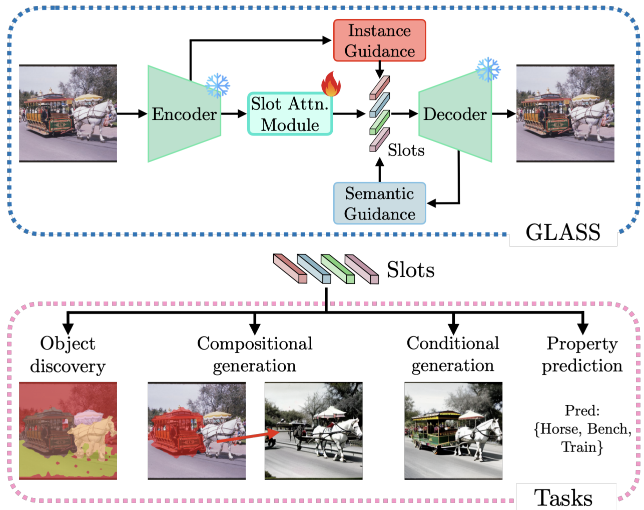

Figure 1. *(top)* **High-level architecture of GLASS.** GLASS employs semantic and instance guidance modules to generate a semantic guidance signal using the decoder network and a instance guidance signal using the encoder features. This helps our method to learn superior slot embeddings for various downstream tasks compared to existing slot attention methods. *(bottom)* **GLASS can perform multiple tasks** using the learned slot embeddings, such as object discovery, compositional generation, conditional generation (reconstruction), and instance-level property prediction.

[*]The code will be published upon the acceptance of the paper.

To overcome the above issues, we propose **G**uided-**La**tent **S**lot Diffu**s**ion (GLASS), a slot attention-based model that uses a pre-trained diffusion decoder for reconstructing the input image and an MLP decoder for reconstructing the encoder features. GLASS relies on two key observations: *(1)* Learning in the space of images generated using diffusion models allows to generalize well to real images because the distribution of the generated images mimics the real data distribution very well [23, 58, 64, 66], and *(2)* learning with generated images allows us to use a pre-trained diffusion model, such as Stable Diffusion [56], as a pseudo ground-truth generation engine. To this end, GLASS relies on a novel semantic guidance module, which uses the diffusion decoder to generate the pseudo-semantic mask. The semantic guidance module helps GLASS solve over-segmentation issues and obtain precise boundaries.

However, using semantic guidance alone biases the slots to semantic classes instead of instances in an image, causing under-segmentation. To resolve this issue, we propose an instance module in the form of an MLP decoder, which reconstructs the encoder features to counteract slots drifting towards semantic classes, and instead guides them to be instance focused. This enables the slots to learn better slot embeddings, which are more instance centric. GLASS's use of semantic and instance guidance modules coupled with a diffusion decoder enables it to faithfully reconstruct / conditionally generate the input image. More importantly, GLASS for the first time enables the compositional generation of complex real-world scenes with slot-attention methods. Fig. 1 illustrates the high-level architecture and the downstream tasks our model supports.

Through our experiments, we show that GLASS outperforms existing SotA OCL methods [32, 36, 60, 71], significantly improving instance-level object discovery (ca. +9% mIoU$_i$ on VOC [21] and +5% mIoU$_i$ on COCO [44]). Our method also outperforms SotA OCL methods on the task of (zero-shot) instance-level segmentation (on Object365 [61] and CLEVRTex [37] datasets). GLASS further establishes a new SotA FID score among OCL methods for conditional image generation tasks and shows that compositional generation is possible with slot attention models for complex real-world scenes. Moreover, we find that our approach surpasses language-based methods [46, 54, 70, 77] for semantic-level object discovery. Finally, we show that GLASS outperforms weakly-supervised variant of a SotA OCL method [32] that rely on extra information like bounding box information or knowing the number of objects in a scene.

2. Related work

Object-centric learning decomposes a multi-object scene into a set of composable and meaningful entities using an autoencoding objective [5, 13, 18, 20, 24, 25, 34, 45, 63]. OCL methods are object-level representation learning approaches that can be employed for various downstream tasks (*cf*. Tab. 1). Among OCL approaches, slot attention methods have proven the most effective; they employ an architectural inductive bias to learn object embeddings, so-called "slots", from the input image. Until recently, a major obstacle for slot attention had been their poor performance on real-world images [74]. This was partially alleviated using large-scale pre-trained models as encoder [60] and decoder [32, 71], which allowed to apply slot attention beyond synthetic imagery. Yet, these models still suffer from the part-whole hierarchy ambiguity, hampering the quality of the learned slot embedding, resulting in poor downstream performance. Our method aims to solve this issue using our proposed semantic and instance guidance modules.

	Semantic-level OD methods						OCL methods				
	DeepSpectral [10]	SegCLIP [46]	CLIPpy [54]	DiffuMask [70]	Dataset Diff. [50]	EmerDiff [49]	Slot Attention [45]	DINOSAUR [60]	SPOT [36]	StableLSD [32]	GLASS (ours)
(1) iOD	✗	✗	✗	✗	✗	✗	(✓)	✓	✓	✓	✓
(2) sOD	✓	✓	✓	✓	✓	✓	(✓)	✓	✓	✓	✓
(3) Latent object file (PP)	✗	✗	✗	✗	✗	✗	(✓)	✓	✓	✓	✓
(4) Cond. Gen. (CG)	✗	✗	✗	✗	✗	✗	(✓)	✗	✗	✓	✓
(5) Comp. Gen. (CPG)	✗	✗	✗	✗	✗	✗	(✓)	✗	✗	(✓)	✓

Table 1. **GLASS's capabilities compared with prior work for solving downstream tasks on *real-world scenes*.** The rows indicate if each method *(1)* can perform instance-level object discovery (OD); *(2)* can perform semantic-level OD; *(3)* provide latents for each object, which enables instance-level property prediction; *(4)* can reconstruct the given image from its latents; and *(5)* can compositionally generate new scenes. (✓): limited performance.

Weakly-supervised object-centric learning. Several works have tried to tackle the part-whole hierarchy ambiguity plaguing OCL with additional weak supervision signals. Video-based OCL methods used motion [41, 65] and depth cues [17], while image-based OCL methods have used position [40] and shape [16] information. Existing weakly-supervised image-based OCL methods [16, 40] remain limited to synthetic datasets, while we focus on complex real-world scenes. GLASS also uses auxiliary information in the form of automatically generated captions. To show the effectiveness of our method, we additionally compare GLASS to a weakly-supervised variant of StableLSD [32] (since this model is closest in capabilities to GLASS, see Tab. 1).

Semantic-level object discovery. Recently, there has been a large interest in using pre-trained features from large-scale foundational models [6, 51, 53, 56] for semantic segmentation. Some of these models [9, 14, 38, 39, 48, 50, 52, 54, 70, 73] rely on language cues like image-level labels or captions to extract features, which are suitable for semantic segmentation. Other methods like [12, 47, 49, 75] do not require any additional information and use clustering or graph cuts with pre-trained features for semantic segmentation. Unlike OCL methods, these approaches are specifically designed to perform semantic-level segmentation, *i.e.* they cannot distinguish between objects of the same class. Also, these methods cannot generate images conditionally or compositionally, nor perform object-level reasoning (see Tab. 1). We compare such methods with a semantic-focused version of GLASS to show its efficacy on semantic-level object discovery.

3. Preliminaries

Slot attention [45] is an iterative refinement scheme based on a set $\mathbf{S} \in \mathbb{R}^{O \times d_{\text{slots}}}$, composed of O slots of dimension d_{slots}, which are initialized either randomly or via a learned function. Once initialized, the representations of the slots are updated iteratively using a GRU network [11] based on the feature matrix $\mathbf{H} \in \mathbb{R}^{N \times d_{\text{input}}}$ of the encoded input image, containing N feature vectors of dimension d_{input}, and the previous state of the slots. Slot attention uses standard dot-product attention [67] for computing the attention matrix $\mathbf{A} \in \mathbb{R}^{N \times O}$, *normalized across slots*. This normalization causes the slots to compete with each other, leading to a meaningful decomposition of the input image. The slots are updated using a weighted combination of the input features $\mathbf{H}$ and the computed attention matrix $\mathbf{A}$. Formally, this can be written as

$$\hat{\mathbf{S}} = \left(\frac{\mathbf{A}_{i,j}}{\sum_{l=1}^{N} \mathbf{A}_{l,j}} \right)_{i,j}^{\top} v(\mathbf{H}) \\ \text{with } \mathbf{A}(\mathbf{S}, \mathbf{H}) = \operatorname{softmax}\left(\frac{k(\mathbf{H}) q(\mathbf{S})^{\top}}{\sqrt{D}} \right), \tag{1}$$

where k, q, and v are learnable linear functions for mapping the slots and input features to the same D dimensions. The updated set of slots $\hat{\mathbf{S}}$ is fed into a decoder model to reconstruct the input. The decoder model can be a simple MLP [69], a transformer [63], or a diffusion model [32, 71]. Slot attention methods are trained using the mean squared error loss between the input and reconstructed input signal.

Latent diffusion models (LDM) [56] learn to generate an image by first iteratively destructing the image by adding Gaussian noise at each time step. This noising process is called the “forward process”. The “reverse process”, or generation step, then involves learning a neural network ϵ_θ that predicts the noise added in each forward diffusion step and removes the noise from the noisy image at each time step. An additional conditioning signal, most commonly in the form of text, is provided to the diffusion model for enabling conditional generation from the diffusion model. The parameters of ϵ_θ are learned by minimizing the mean squared error between the predicted and ground-truth noise added for each time step in the denoising process. Once trained, an image can be generated by sampling a random noise vector and running the reverse process with a given conditional signal. The most common choice for the denoising network ϵ_θ is a U-Net [57] with layers of self- and cross-attention at multiple resolutions. The cross-attention layers cross-attend between the conditioning signal and the pixel features.

4. Guided Latent Slot Diffusion (GLASS)

GLASS is based on training a slot attention module on the features of a DINOv2 [51] (encoder) model and uses a pre-trained Stable Diffusion (SD) model [56] (decoder) to reconstruct the image, as well as a small MLP model to reconstruct the encoder features. GLASS leverages the diffusion decoder and a pre-trained caption generation model [43] to create a guidance signal (segmentation masks) to guide slots.

A key design choice in our proposed method is to learn the slot attention module in the space of generated images from a pre-trained diffusion model. This enables us to use the cross-attention layers in the U-Net [57] of the diffusion decoder for obtaining the semantic mask for the given image. Let us now describe each step in detail.

Conditional image generation. Given an input image $\mathbf{I}_{\text{inp}}$, we first pass it through a caption generator (BLIP-2 [43]) to generate a caption $\mathcal{P}_{\text{cap}}$ that describes the input image. We extract the nouns from the generated caption using a part-of-speech (POS) tagger [3] and retain those nouns, $\mathcal{C} = \{\mathrm{c}_1, \mathrm{c}_2, \ldots, \mathrm{c}_k\}$, that belong to the set of COCO's class labels. We then create a prompt, $\mathcal{P} = [\mathcal{P}_{\text{cap}}; \mathcal{C}]$, by concatenating the generated caption and the extracted class labels. This prompt $\mathcal{P}$ is fed into a text embedder, here CLIP [53], to obtain an embedding $\mathbf{Y} \in \mathbb{R}^{U \times d_{\text{token}}}$, where U is the number of tokens of dimension d_{token}. We then generate an image $\mathbf{I}_{\text{gen}}$ by sampling random noise from $\mathcal{N}(0, I)$ and running the “reverse process” on a pre-trained diffusion model with $\mathbf{Y}$ as a conditioning signal.

Pseudo ground-truth generation module. For extracting the cross-attention map at time t for layer l from the diffusion model, we create a new prompt consisting of a single token, namely one of the class tokens from $\mathcal{C}$. The cross-attention map for the target label can be computed using standard dot-product attention between the linear projections of the ground-truth class label embedding and the noisy image features in a common d-dimensional space. This is done for each target class label in $\mathcal{C}$. The final cross-attention map $\mathbf{A}_{\text{CA}} \in [0,1]^{H \times W \times C}$ is obtained by resizing and averaging the extracted cross-attention maps across different time steps and resolutions. Here, H and W are the sizes of the input embedding and C is the number of target classes. The obtained cross-attention maps are often noisy and require further refinement. Recently, several works have addressed the problem of refining such cross-attention maps [39, 50, 70]. We follow [50] and use self-attention maps for

CVPR #12908

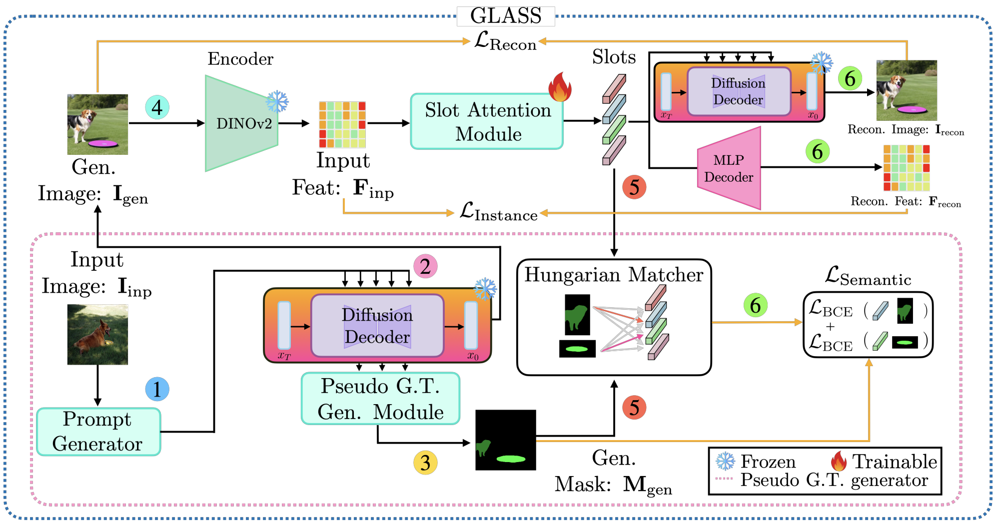

Figure 2. **Network architecture of GLASS.** ① The input image $\mathbf{I}_{\text{inp}}$ is fed to a prompt generator for generating a prompt $\mathcal{P}$, which is obtained by concatenating the generated caption $\mathcal{P}_{\text{cap}}$ and the extracted class labels from $\mathcal{P}_{\text{cap}}$. ② A random noise vector, along with the generated prompt $\mathcal{P}$, is used to generate an image $\mathbf{I}_{\text{gen}}$ using a pre-trained diffusion decoder module. ③ The cross-attention layers of the diffusion model, along with self-attention layers, are used in the pseudo ground-truth generation module to generate the semantic mask $\mathbf{M}_{\text{gen}}$ for $\mathbf{I}_{\text{gen}}$. ④ The generated image is passed through an encoder model (DINOv2) followed by a slot attention module to generate slots. ⑤ The slots are matched with their corresponding object masks from $\mathbf{M}_{\text{gen}}$ using the Hungarian matcher module. ⑥ The slot attention module is trained end-to-end using the mean squared error ($\mathcal{L}_{\text{Recon}}$) between the reconstructed ($\mathbf{I}_{\text{recon}}$) and the generated ($\mathbf{I}_{\text{gen}}$) image, and our *semantic* ($\mathcal{L}_{\text{Semantic}}$) and *instance* ($\mathcal{L}_{\text{Instance}}$) guidance losses. GLASS is trained on generated images only; the real image is used for prompt generation.

refining the cross-attention maps. In particular, the refined mask $\mathbf{M}_{\text{ref}}$ is obtained by exponentiating the self-attention map $\mathbf{A}_{\text{SA}} \in [0,1]^{H\times W\times H\times W}$ and multiplying with the cross-attention map $\mathbf{A}_{\text{CA}}$ as described in [50]. The final semantic mask $\mathbf{M}_{\text{gen}}$ is obtained by taking the pixel-wise $\arg\max$ of $\mathbf{M}_{\text{ref}}$ for all target class labels in $\mathcal{C}$ to find which class is responsible for a given pixel. Finally, a range-based thresholding is used to classify each pixel as foreground or background. See supplemental for details.

Slot matching. Once the images $\mathbf{I}_{\text{gen}}$ and their corresponding pseudo ground-truth semantic masks $\mathbf{M}_{\text{gen}}$ are generated, we can use these semantic masks to guide the slots. First, we pass the generated image $\mathbf{I}_{\text{gen}}$ through the encoder and the slot attention module to obtain a slot decomposition. We extract the predicted masks for each slot using the attention matrix $\mathbf{A}(\mathbf{S},\mathbf{H})$ from Eq. (1) and resize them to the resolution of the generated semantic mask $\mathbf{M}_{\text{gen}}$. We then assign each predicted mask to the components of the generated semantic masks. This akin to solving a bipartite matching problem for which we use Hungarian matching [42]. Formally, given O slots with their predicted masks and a semantic mask containing F segments, the binary matching matrix $\mathbf{P} \in \{0,1\}^{O\times F}$ can be computed using the Hungarian algorithm that minimizes the cost $c_{i,j}$ of assigning slot o_i to segment m_j in the generated mask $\mathbf{M}_{\text{gen}}$:

$$\min_{\mathbf{P}} \sum_{i=1}^{O}\sum_{j=1}^{F} -c_{i,j}p_{i,j}, \tag{2}$$

where $p_{i,j} \in \{0,1\}$ indicates whether o_i is matched with segment m_j. The optimization is constrained to assign each slot to one and only one segment. The cost $c_{i,j}$ is calculated using the mean Intersection over Union (IoU) between the predicted mask of slot o_i and segment m_j of the generated semantic mask $\mathbf{M}_{\text{gen}}$. The optimal assignment is the one that maximizes the overall mean IoU.

Loss function. Once the assignment is complete, our guided slot attention model is trained end-to-end using the mean squared error loss ($\mathcal{L}_{\text{MSE}}$) between the generated image $\mathbf{I}_{\text{gen}}$ and reconstructed image $\mathbf{I}_{\text{recon}}$, as well as our (semantic) *guidance loss*, *i.e.* a binary cross-entropy loss ($\mathcal{L}_{\text{BCE}}$) between $\mathbf{M}_{\text{gen}}$ and the predicted mask from the slots $\mathbf{A}(\mathbf{S},\mathbf{H})$. The binary cross-entropy loss is only computed on the matched slots, according to the matching matrix $\mathbf{P} = \mathbf{P}(\mathbf{M}_{\text{gen}}, \mathbf{A}(\mathbf{S},\mathbf{H}))$. Simply using the image reconstruction and semantic guidance loss would lead the slot representation to drift towards semantic classes and not to objects. We tackle this semantic drift problem by adding a feature reconstruction loss, which we term instance guidance loss. The instance guidance loss is given by the mean squared error between the input ($\mathbf{F}_{\text{inp}}$) and reconstructed ($\mathbf{F}_{\text{recon}}$) features (see Fig. 2). Our full loss is given by

$$\begin{aligned}\mathcal{L} = \underbrace{\mathcal{L}_{\text{MSE}}(\mathbf{I}_{\text{gen}}, \mathbf{I}_{\text{recon}})}_{\text{Recon. Loss } (\mathcal{L}_{\text{Recon}})} + \lambda_s \underbrace{\mathcal{L}_{\text{BCE}}(\mathbf{P}(\mathbf{M}_{\text{gen}}, \mathbf{A}(\mathbf{S},\mathbf{H})))}_{\text{Semantic Guidance } (\mathcal{L}_{\text{semantic}})} \\ + \lambda_i \underbrace{\mathcal{L}_{\text{MSE}}(\mathbf{F}_{\text{inp}}, \mathbf{F}_{\text{recon}})}_{\text{Instance Guidance } (\mathcal{L}_{\text{instance}})} .\end{aligned} \tag{3}$$

CVPR
#12908

The semantic guidance loss helps learn a slot representation that adheres to object boundaries and does not split the object into multiple slots (*i.e*., avoids over-segmentation) but causes the slots to focus on semantics and not on instances. The feature reconstruction loss helps with the semantic drift problem as features from a pre-trained ViT model already exhibit instance-aware properties [19], but using them without semantic guidance results in over- and under-segmentation issues. Thus, when instance and semantic guidance are coupled, the slots are bound to the instances instead of semantics and avoid the part-whole ambiguity, *cf*. also Fig. 4.

We separate the training process of GLASS into two phases: In phase-1, only the slot attention module and MLP decoder are trained. This helps in learning slot embeddings that bind to instances. In phase-2, we jointly train both the slot attention module with diffusion and MLP decoders. In this phase, we use a small learning rate for the slot attention and MLP decoder modules and a higher learning rate for the diffusion decoder. The second phase helps the diffusion decoder align to slot embeddings and produce high-fidelity images. Unless otherwise stated, we use $\lambda_s = 0.7$ and $\lambda_i = 0.9$ for all our experiments. Fig. 2 shows our full architecture and illustrates each step. Further details about the training and datasets are provided in the supplemental.

5. Experiments

The main focus of our work is to learn better representations of objects, *i.e*. slot embeddings. To assess the effectiveness of the learned representation, we test GLASS on various tasks such as object discovery, instance-level property prediction, reconstruction, and compositional generation. Our method uses generated captions from BLIPv2 [43], which is trained on image-caption pairs mined from the web; thus, our model can be considered very weakly (coincidentally) supervised. Therefore, we test it against other weakly-supervised OCL methods. We also propose a variant of GLASS termed GLASS†, which uses ground-truth class labels associated with the input image instead of the generated caption to generate and extract the semantic guidance signal.

5.1. Instance-aware object discovery

The standard way to test how well the slots bind to an object is to evaluate on the object discovery task, *i.e*., producing a set of masks that cover the independent objects appearing in an image. We compare GLASS against existing SotA object-centric methods using the standard multi-object discovery metrics popular in the OCL literature [32, 60, 71]. This includes *(i)* the mean Intersection over Union between the predicted masks from the slots, which are computed using the attention weights $\mathbf{A}(\mathbf{S}, \mathbf{H})$ as defined in Eq. (1), and the ground-truth *instance* masks, mIoU_i, *(ii)* the mean Best Overlap over instance-level masks, mBO_i, and *(iii)* over class-level masks, mBO_c. Please see the supplemental for details and additional results for the foreground adjusted rand index, FG-ARI. Tab. 2 shows that GLASS outperforms all previous OCL methods across mIoU_i, mBO_i, and mBO_c metrics by a wide margin. Fig. 3 shows qualitative results for object discovery compared to DINOSAUR [60], StableLSD [32], and SPOT [36]. They show that our method decomposes a scene in a more instance-centric way with sharper boundaries, no object splitting, and cleaner background segmentation. Importantly, unlike SPOT, our model can correctly segment different instances of the same class of objects, see Fig. 3.

Model	COCO (in %, all ↑)			VOC (in %, all ↑)		
	mIoU_i	mBO_i	mBO_c	mIoU_i	mBO_i	mBO_c
SA* [45] NeurIPS'20	–	17.2	19.2	–	24.6	24.9
SLATE* [62] ICLR'22	–	29.1	33.6	–	35.9	41.5
DINOSAUR-MLP [60] ICLR'23	26.8	28.1	32.1	39.1	39.7	41.2
DINOSAUR-Trans. [60] ICLR'23	31.6	33.3	41.2	42.0	43.2	47.8
SPOT [36] CVPR'24	34.0	35.0	44.7	48.8	48.3	55.6
SlotDiffusion* [71] NeurIPS'23	–	31.0	35.0	–	50.4	55.3
StableLSD [32] NeurIPS'23	24.7	25.9	30.0	30.0	30.4	33.1
GLASS† (ours)	**39.0** (+5.0)	**40.8** (+5.8)	**48.7** (+4.0)	57.8 (+9.0)	58.5 (+8.1)	61.5 (+5.9)
GLASS (ours)	38.9 (+4.9)	40.6 (+5.6)	48.5 (+3.8)	**58.1** (+9.3)	**58.9** (+8.5)	**62.2** (+6.6)

Table 2. **Comparison between OCL methods for instance-aware object discovery.** GLASS and GLASS† clearly outperform all other SotA OCL methods on the multi-object discovery metrics. The best value is highlighted in **bold**, the second best is underlined. * numbers are taken from [36]. Values in parentheses denote the improvement of GLASS over the previous SotA method. Tab. 9 shows additional info. about the methods *e.g*. pre-trained models used, input modalities, and downstream capabilities for each method.

Model	COCO (in %)			VOC (in %)		
	SO (↑)	PO (↓)	GO (↓)	SO (↑)	PO (↓)	GO (↓)
StableLSD [32] NeurIPS'23	10.2	87.3	1.6	6.7	91.6	0.40
DINOSAUR [60] ICLR'23	22.1	71.2	2.1	22.4	70.2	0.07
SPOT [36] CVPR'24	24.7	69.7	0.01	26.2	65.0	**0.00**
GLASS†	**27.3**	49.6	0.01	**30.4**	**26.2**	0.67
GLASS	25.2	**45.9**	**0.00**	26.7	42.3	0.01

Table 3. **SO-PO-GO metrics.** Our method has a higher % of slots that bind to single object compared to baselines, while also being less prone to over-segmentation and under-segmentation as seen by PO and GO metrics. % of slots binding to background not shown.

SO-PO-GO metrics. A major reason for our method's success is because it reduces the over- and under-segmentation (part-whole ambiguity) issues, which plague existing OCL methods. To quantify this further, we evaluate the effectiveness of GLASS in resolving these ambiguities using the SO-PO-GO metric proposed by [22]. The metric reports the percentage of slots that bind to a single object (*SO*), slots that bind to part of an object (*PO*), and slots that bind to a group of objects (*GO*). As seen in Tab. 3, our method has a much higher percentage of slots that bind to a single object while reducing the number of slots that bind to parts of objects compared to SotA OCL methods.

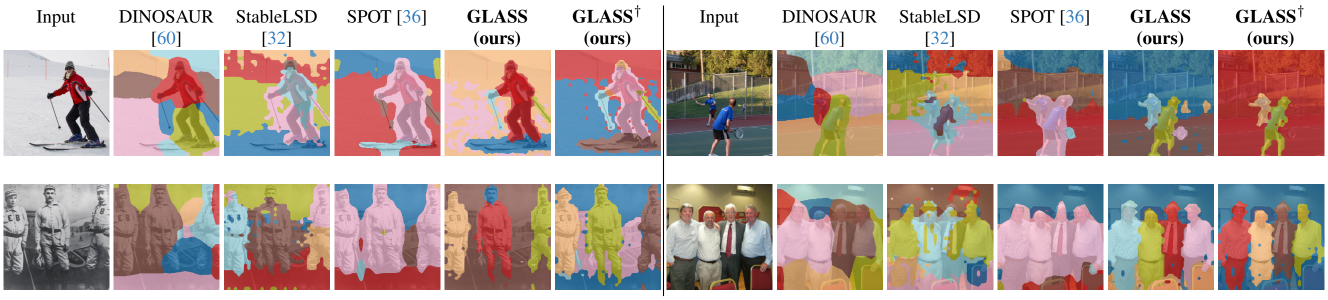

Figure 3. **Qualitative comparison for object discovery.** GLASS and GLASS† can decompose an image at an instance level and reduces over and under-segmentation of objects. Our method also yields cleaner boundaries for objects compared to StableLSD and DINOSAUR.

Model	CLEVRTex (in %, all ↑)		Obj365 (in %, all ↑)	
	mIoU$_i$	mBO$_i$	mIoU$_i$	mBO$_i$
StableLSD [32] NeurIPS'23	24.0	27.6	14.8	16.9
DINOSAUR [60] ICLR'23	30.2	35.1	16.2	18.9
SPOT [36] CVPR'24	39.5	43.7	18.0	20.7
GLASS†	**47.2**	**52.3**	**19.6**	**22.4**
GLASS	46.1	50.1	18.6	21.4

Table 4. **Zero-shot object discovery**. Our method outperforms the baseline methods on the task of zero-shot object discovery.

Model	VOC (in %, all ↑)		
	mIoU$_i$	mBO$_i$	mBO$_c$
StableLSD [32] NeurIPS'23	30.0	30.4	33.1
StableLSD-Bbox	30.5	37.8	42.2
StableLSD-Dynamic	30.8	38.2	43.4
GLASS† (ours)	57.8	**58.5**	61.5
GLASS (ours)	**58.1**	58.9	**62.2**

Table 5. **Comparison with weakly-supervised baselines,** *i.e.* variants of StableLSD. GLASS clearly outperforms the weakly-supervised variants of the StableLSD model even though it uses weaker supervision than these variants.

Zero-shot learning. We next show that resolving the part-whole ambiguity also helps object discovery (OD) in a zero-shot manner. As the slots are now biased towards objects, they can better segment scenes compared to baseline methods even when not trained on them. We take GLASS trained on the COCO dataset and report the zero-shot OD results on the CLEVRTex [37] and Obj365 [61] datasets, see Tab. 4. We obtained the masks for the Obj365 dataset by prompting SAMv2 [55] with ground-truth bounding boxes. We observe that our approach again outperforms SotA OCL methods.

Comparison to weakly-supervised OCL. Our method can be considered (very) weakly supervised due to its dependence on the BLIP-2 [43] model for caption generation. We show that this form of (very) weak supervision performs much better than using more expensive weakly supervised signals, such as bounding boxes or knowing the number of objects in the scene. In particular, we compare our method against two weakly-supervised variants of StableLSD: *(i)* StableLSD-BBox, which uses the bounding-box information associated with each object for initializing the slots. This form of guidance has been previously used in [41]. *(ii)* StableLSD-Dynamic, which, instead of having a fixed number of slots for each scene, dynamically assigns each scene the number of slots equal to the number of objects present. This technique was useful for addressing the issue of part-whole ambiguity, leading to better object discovery [78]. We choose StableLSD for comparison since it is closest to our model regarding the downstream tasks it can perform (see Tab. 1). As seen in Tab. 5, the weakly-supervised variants of StableLSD outperform StableLSD. *Importantly*, GLASS outperforms both weakly-supervised methods even though it uses a weaker supervision signal.

Importance of semantic and instance guidance. Next, we evaluate the contribution of the semantic and instance guidance losses. Tab. 6a shows the mIoU$_i$ metrics with different combinations of our three loss functions ($\mathcal{L}_{\text{Recon}}$, $\mathcal{L}_{\text{Semantic}}$, and $\mathcal{L}_{\text{Instance}}$). We observe that combining semantic and instance losses together produces much better results than using them individually. More importantly, the qualitative results in Fig. 4 show that using only the reconstruction loss results in a noisy segmentation (over- and under-segmentation). Adding the semantic loss helps in obtaining more precise boundaries, making the segmentation much less noisy. However, just using the semantic loss causes semantic drift and binds slots to semantic classes (under-segmentation); adding the instance guidance breaks the semantic drift problem and makes slots bind to objects instead of semantic classes. Thus, utilizing both semantic and instance guidance alleviates the over- and under-segmentation issue, making the learned slot embeddings more powerful for downstream tasks.

Performance with different encoder networks. We next ablate the dependence of GLASS on the encoder architecture. We benchmark the performance for three different encoder models, namely Masked Auto Encoders (MAE) [27], DINOv2 [51], and DINOv1 [8]. As seen in Tab. 6b, our method is robust to the choice of the encoder model. Moreover, it outperforms the model closest to our method regarding downstream capabilities (StableLSD) for all encoder model architectures.

CVPR #12908

Loss term	mIoU$_i$ (in %, ↑) COCO	VOC
$\mathcal{L}_{\text{Recon}}$	30.0	34.4
$\mathcal{L}_{\text{Recon}} + 0.7\mathcal{L}_{\text{Semantic}}$	30.9	55.1
$\mathcal{L}_{\text{Recon}} + 0.9\mathcal{L}_{\text{Instance}}$	29.3	38.9
$\mathcal{L}_{\text{Recon}} + 0.7\mathcal{L}_{\text{Semantic}} + 0.9\mathcal{L}_{\text{Instance}}$	**38.9**	**58.1**

(a) **Importance of semantic and instance guidance losses**. A combination of semantic and instance loss terms performs the best.

Encoder	mIoU$_i$ (in %, ↑) COCO	VOC
Baseline (StableLSD [32])	24.7	30.0
GLASS w/ MAE [27]	30.0	41.1
GLASS w/ DINOv1 [7]	31.4	54.2
GLASS w/ DINOv2 [51]	**38.9**	**58.1**

(b) **Effect of encoder network.** GLASS is robust to the encoder architecture and outperforms the baseline even with weaker encoder networks.

Model	mIoU$_i$ (in %, ↑) COCO	VOC
GLASS w/ SAMv2	30.4	42.1
GLASS†	31.1	**58.6**
GLASS	**32.2**	55.6

(c) **Effectiveness of pseudo GT semantic mask**. Using masks from the decoder performs better than masks obtained from SAMv2.

Table 6. **Ablation study.** *(a)* We study the impact of different loss terms on GLASS, *(b)* the impact of different encoder architectures, and *(c)* the impact of using different guidance generation on the performance of our approach on the instance-level object discovery task.

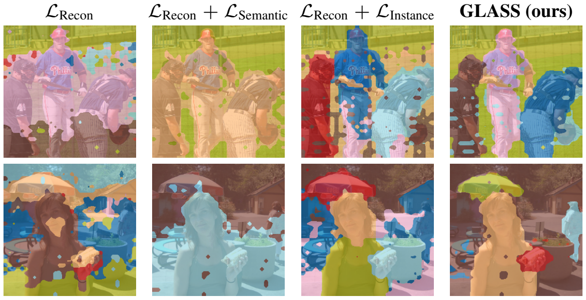

Figure 4. **Qualitative results showing the importance of joint semantic and instance guidance.** Using both guidances together provides precise boundaries and biases the slots to object instances.

Figure 5. **Qualitative comparison for conditional image generation.** GLASS and GLASS† reconstruct the input scene more faithfully with a high degree of detail as compared to StableLSD.

Model	PSNR (in dB, ↑)	SSIM (↑)	LPIPS (↓)	FID (↓)
StableLSD [32] NeurIPS'23	10.92	0.20	0.72	140.62
GLASS† (ours)	10.88	0.20	**0.59**	79.61
GLASS (ours)	**10.93**	**0.21**	**0.59**	**71.30**

Table 7. **Conditional generation.** Comparison between StableLSD and our approach for the conditional generation / recon. task.

Model	VOC		COCO	
	Acc (in %, ↑)	MSE (↓)	Acc (in %, ↑)	MSE (↓)
StableLSD [32] NeurIPS'23	55.1	0.039	16.4	0.062
GLASS (ours)	**58.1**	**0.037**	**20.8**	**0.059**

Table 8. **Instance-level property prediction.** Comparison between StableLSD and GLASS for the property prediction task.

Importance of pseudo ground-truth generation module. A key advantage of our method is utilizing the decoder model for both decoding the slots and also as semantic guidance generator, resulting in no additional dependency for guidance generation. We next show that our method of obtaining the guidance signal is superior to obtaining the guidance signal from models such as SAMv2 [55]. To assess the impact of the semantic guidance signal, we set λ_i or the instance guidance loss to zero and $\lambda_s = 1$ for this experiment. As seen in Tab. 6c, our pseudo-ground truth signals lead to better performance of our method over using masks from SAMv2. This is because, without prompting, SAMv2 produces masks that are either over- or under-segmented compared to masks obtained with our method. To use SAMv2 effectively, we need an additional prompt, *e.g*. a bounding box, but this form of supervision is more expensive than generated captions or image-level labels.

5.2. Generative capabilities

Conditional generation/reconstruction. Using a diffusion-based decoder in GLASS enables our model to conditionally generate the input image back from the slots and, more importantly, to be able to compositionally generate new scenes. We benchmark GLASS against StableLSD for conditional image generation, as this is the only OCL model to date to be able to reconstruct complex real-world images. We report the PSNR, SSIM [68], LPIPS [76], and FID [28] metrics. Both quantitatively (see Tab. 7) and qualitatively (*cf*. Fig. 5), our method outperforms StableLSD. The qualitative results show that GLASS can reconstruct the input image more faithfully and with higher fidelity.

Compositional generation. To the best of our knowledge, GLASS is the first among slot attention-based methods to be able to compositionally generate complex real-world scenes with high fidelity. We show examples where objects can be removed from an input scene by removing a slot, or objects can be added to a scene by adding the slots from a different scene. We show qualitative results in Fig. 6. Compositional generation with StableLSD results in very low-fidelity images, see supplemental for details.

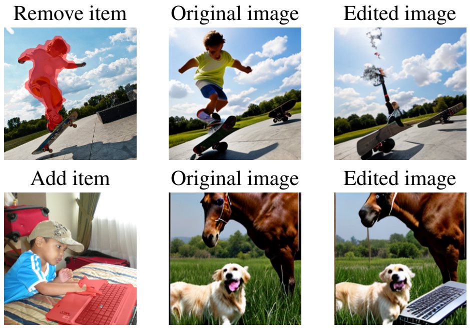

Figure 6. **Compositional generation.** GLASS enables compositional image generation of real-world complex scenes. Here, the masked object *(in red)* is the slot to be removed from or added to the original image. Please see supplemental for more results.

5.3. Property prediction

Instance-level property prediction assesses the quality of the slot representation. In this task, we predict object properties, such as class labels and object positions (centre of the object's bounding box) in the input images from the learned slot embeddings. We compare the informativeness of the features learned by slots of GLASS and StableLSD. We report top-1 accuracy for label prediction and mean squared error for predicting the object's center. As seen in Tab. 8, GLASS consistently outperforms StableLSD for both tasks, indicating that our learned slots contain more informative features about the object than StableLSD's slot embeddings.

5.4. Semantic-level object discovery

Since our method makes use of large-scale pre-trained foundational models [7, 51, 56], we also compare it against other approaches [*e.g.*, 12, 14, 46, 70] utilizing the features from these foundational models. However, these models are *only* able to perform semantic-level segmentation. Our method is designed for instance-level segmentation but can also be modified to enable semantic-level segmentation. We show results for a special case of our model (semantic-focused GLASS) where we purposefully make our model under-segment the image (one slot is responsible for multiple objects belonging to the same class). For this, we set the instance guidance loss term to a low value ($\lambda_i = 0.1$) during training. For this task, we report the mIoU$_c$ metric computed between the predicted masks from the slots and the ground-truth *semantic* masks.

Tab. 9 shows that our method outperforms not only all object-centric learning methods but also methods that rely on features from large-scale models for performing semantic-level object discovery. We attribute the improvement of GLASS over other methods that use foundational models to its careful interplay of features between the different foundational models: Our approach aggregates features from a foundation model (DINOv2 [51]) but this feature aggregation is guided by our semantic guidance module, which helps it achieve precise boundaries. This interpretation is supported by the observation that GLASS outperforms models such as Dataset Diffusion [50] even though, just like GLASS, they use Stable Diffusion features for creating pseudo masks.

Model	Downstr. tasks	Input	Pre-trained models	mIoU$_c$ (in %, ↑)	
				COCO	VOC
MaskCLIP [77] ECCV'22	sOD	$\mathcal{I} + \mathcal{C}$	CLIP	20.6	38.8
SegCLIP [46] ICML'23	sOD	$\mathcal{I} + \mathcal{C}$	CLIP	26.5	52.6
CLIPPy [54] ICCV'23	sOD	$\mathcal{I} + \mathcal{C}$	CLIP	32.0	52.2
OVSeg [72] CVPR'23	sOD	$\mathcal{I} + \mathcal{C}$	CLIP	25.1	53.8
DeepSpectral [47] CVPR'22	sOD	$\mathcal{I}$	DINO	–	37.2
COMUS [75] ICLR'23	sOD	$\mathcal{I}$	DINO	–	50.0
DiffuMask [70] NeurIPS'23	sOD	$\mathcal{I} + \mathcal{C}$	SD + CLIP + [1]	–	57.4
Dataset Diffusion [50] NeurIPS'23	sOD	$\mathcal{I}$	SD + BLIP-2	34.2	64.8
DiffCut [12] NeurIPS'24	sOD	$\mathcal{I}$	SD	34.1	65.2
OVDiff [38] ECCV'24	sOD	$\mathcal{I} + \mathcal{C}$	SD + DINO + CLIP	34.6	<u>66.3</u>
EmerDiff [49] ICLR'24	sOD	$\mathcal{I}$	SD	33.1	40.3
DINOSAUR-MLP [60] ICLR'23	iOD + PP	$\mathcal{I}$	DINO	31.7	41.0
DINOSAUR-Transformer [60] ICLR'23	iOD + PP	$\mathcal{I}$	DINO	40.6	47.5
SPOT [36] CVPR'24	iOD + PP	$\mathcal{I}$	DINO	<u>44.6</u>	55.3
StableLSD [32] NeurIPS'23	OD + PP + CG	$\mathcal{I}$	SD + DINOv2	29.5	32.9
GLASS (ours)	iOD + PP + CG + CPG	$\mathcal{I}$	SD + DINOv2 + BLIP-2	**46.7** (+2.1)	**68.9** (+2.6)

Table 9. **Comparison on semantic-level object discovery** We compare our method with baselines that use features from foundational models for semantic-level object discovery. We divide the baselines into training-based *(top)*, training-free *(middle)*, and OCL methods *(bottom)*. *Downstream tasks* denote a model's capability of solving the following tasks: iOD /sOD – instance-/semantic-level object discovery, PP – instance-level property prediction, CG – conditional generation, and CPG – compositional generation. *Input* denotes the input signal the model itself trains on, where $\mathcal{I}$ – image, $\mathcal{C}$ – captions, and $\mathcal{L}$ – image-level labels. *Pre-trained models* denote the underlying frozen foundation models used in the method. **Note:** *GLASS is not designed for sOD, but it is controllable and can be tuned explicitly for either sOD or iOD task.*

6. Conclusion

We present GLASS, a novel object-centric learning method that learns in the space of generated images from a pre-trained diffusion model. Our method makes use of semantic and instance guidance in order to learn better instance-centric representations. We clearly outperform previous SotA OCL methods on various tasks: instance-level (zero-shot) object discovery and conditional image generation. Our work also surpasses SotA models that use large-scale pre-trained models for semantic-level object discovery, and learns better slot representation for instance-level property prediction than similarly versatile OCL methods. Notably, our method enables the first application of compositional generation of complex real-world scenes among OCL methods.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, pages 4981–4990, 2018. 8

[2] Rim Assouel, Pau Rodriguez, Perouz Taslakian, David Vazquez, and Yoshua Bengio. Object-centric compositional imagination for visual abstract reasoning. In *ICLR Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022. 1

[3] Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 2004. 3

[4] Ondrej Biza, Robert Platt, Jan-Willem van de Meent, Lawson L.S. Wong, and Thomas Kipf. Binding actions to objects in world models. In *ICLR Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022. 1

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

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 2

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, pages 9912–9924, 2020. 1, 7, 8

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 6

[9] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *CVPR*, pages 11165–11174, 2023. 2

[10] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *CVPR*, pages 16794–16804, 2021. 2

[11] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP*, 2014. 3

[12] Paul Couairon, Mustafa Shukor, Jean-Emmanuel Haugeard, Matthieu Cord, and Nicolas Thome. Zero-shot image segmentation via recursive normalized cut on diffusion features. In *NeurIPS*, 2024. 2, 8

[13] Eric Crawford and Joelle Pineau. Spatially invariant unsupervised object detection with convolutional neural networks. In *AAAI*, pages 3412–3420, 2019. 2

[14] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with MaskCLIP. In *ICML*, 2023. 2, 8

[15] Andrea Dittadi, Samuele Papa, Michele De Vita, Bernhard Schölkopf, Ole Winther, and Francesco Locatello. Generalization and robustness implications in object-centric learning. In *ICML*, pages 5221–5285, 2021. 1

[16] Cathrin Elich, Martin R. Oswald, Marc Pollefeys, and Joerg Stueckler. Weakly supervised learning of multi-object 3D scene decompositions using deep shape priors. *Comput. Vis. Image Und.*, page 103440, 2022. 1, 2

[17] Gamaleldin F. Elsayed, Aravindh Mahendran, Sjoerd van Steenkiste, Klaus Greff, Michael C. Mozer, and Thomas Kipf. SAVi++: Towards end-to-end object-centric learning from real-world videos. In *NeurIPS*, pages 28940–28954, 2022. 2

[18] Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner. GENESIS: Generative scene inference and sampling with object-centric latent representations. In *ICLR*, 2020. 2

[19] Paul Engstler, Luke Melas-Kyriazi, Christian Rupprecht, and Iro Laina. Understanding self-supervised features for learning unsupervised instance segmentation. *arXiv:2311.14665 [cs.CV]*, 2023. 5

[20] S.M. Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E. Hinton. Attend, infer, repeat: Fast scene understanding with generative models. *NIPS*, 2016. 2

[21] Mark Everingham, Luc Van Gool, Christopher K.I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vision*, pages 303–338, 2010. 1, 2

[22] Ke Fan, Zechen Bai, Tianjun Xiao, et al. Unsupervised open-vocabulary object localization in videos. In *ICCV*, pages 13747–13755, 2023. 5

[23] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. In *CVPR*, pages 7382–7392, 2024. 2

[24] 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 *ICML*, pages 2424–2433, 2019. 2

[25] Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv:2012.05208 [cs.NE]*, 2020. 1, 2

[26] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, et al. Kubric: A scalable dataset generator. In *CVPR*, pages 3749–3761, 2022. 1

[27] Kaiming He, Xinlei Chen, Saining Xie, et al. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 6, 7

[28] 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. *NIPS*, 30, 2017. 7

[29] Geoffrey E. Hinton. Some demonstrations of the effects of structural descriptions in mental imagery. *Cognitive Science*, pages 231–250, 1979. 1

[30] Geoffrey E. Hinton. Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence*, 46:47–75, 1990.

[31] Geoffrey E. Hinton. How to represent part-whole hierarchies in a neural network. *Neural Computation*, 35:413–452, 2023. 1

[32] Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. In *NeurIPS*, 2023. 1, 2, 3, 5, 6, 7, 8

[33] 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 *CVPR*, pages 2901–2910, 2017. 1

[34] Rishabh Kabra, Daniel Zoran, Goker Erdogan, Loic Matthey, Antonia Creswell, Matt Botvinick, Alexander Lerchner, and Chris Burgess. SIMONe: View-invariant, temporally-abstracted object representations via unsupervised video decomposition. *NeurIPS*, pages 20146–20159, 2021. 2

[35] Daniel Kahneman, Anne Treisman, and Brian J. Gibbs. The reviewing of object files: Object-specific integration of information. *Cognitive psychology*, pages 175–219, 1992. 1

[36] Ioannis Kakogeorgiou, Spyros Gidaris, Konstantinos Karantzalos, and Nikos Komodakis. SPOT: Self-training with patch-order permutation for object-centric learning with autoregressive transformers. In *CVPR*, pages 22776–22786, 2024. 1, 2, 5, 6, 8

[37] Laurynas Karazija, Iro Laina, and Christian Rupprecht. ClevrTex: A texture-rich benchmark for unsupervised multi-object segmentation. In *NeurIPS Datasets and Benchmarks Track*, 2021. 1, 2, 6

[38] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation. In *ECCV*, 2024. 2, 8

[39] Aliasghar Khani, Saeid Asgari Taghanaki, Aditya Sanghi, Ali Mahdavi Amiri, and Ghassan Hamarneh. SLiMe: Segment like me. In *ICLR*, 2024. 2, 3

[40] Dongwon Kim, Namyup Kim, Cuiling Lan, and Suha Kwak. Shatter and Gather: Learning referring image segmentation with text supervision. In *ICCV*, pages 15547–15557, 2023. 2

[41] Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. In *ICLR*, 2022. 2, 6

[42] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, pages 83–97, 1955. 4

[43] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 3, 5, 6

[44] 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 *ECCV*, pages 740–755, 2014. 1, 2

[45] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *NeurIPS*, pages 11525–11538, 2020. 1, 2, 3, 5

[46] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. SegCLIP: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *ICML*, pages 23033–23044, 2023. 2, 8

[47] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *CVPR*, pages 8364–8375, 2022. 2, 8

[48] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip H.S. Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *CVPR*, pages 19413–19423, 2023. 2

[49] Koichi Namekata, Amirmojtaba Sabour, Sanja Fidler, and Seung Wook Kim. EmerDiff: Emerging pixel-level semantic knowledge in diffusion models. *ICLR*, 2024. 2, 8

[50] Quang Ho Nguyen, Truong Tuan Vu, Anh Tuan Tran, and Khoi Nguyen. Dataset Diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. In *NeurIPS*, 2023. 2, 3, 4, 8

[51] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv:2304.07193 [cs.CV]*, 2023. 1, 2, 3, 6, 7, 8

[52] Koutilya Pnvr, Bharat Singh, Pallabi Ghosh, Behjat Siddiquie, and David Jacobs. LD-ZNet: A latent diffusion approach for text-based image segmentation. In *ICCV*, pages 4157–4168, 2023. 2

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2, 3

[54] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. In *ICCV*, pages 5571–5584, 2023. 2, 8

[55] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, et al. SAM 2: Segment anything in images and videos. *arXiv:2408.00714 [cs.CV]*, 2024. 6, 7

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 3, 8

[57] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 3

[58] Mert Bülent Sarıyıldız, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic ImageNet clones. In *CVPR*, pages 8011–8021, 2023. 2

[59] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, pages 612–634, 2021. 1

[60] Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, et al. Bridging the gap to real-world object-centric learning. In *ICLR*, 2022. 1, 2, 5, 6, 8

[61] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pages 8430–8439, 2019. 2, 6

[62] Gautam Singh, Fei Deng, and Sungjin Ahn. Illiterate DALL-E learns to compose. In *ICLR*, 2021. 1, 5

[63] Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. *NeurIPS*, pages 18181–18196, 2022. 1, 2, 3

[64] Krishnakant Singh, Thanush Navaratnam, Jannik Holmer, Simone Schaub-Meyer, and Stefan Roth. Is synthetic data all we need? Benchmarking the robustness of models trained with synthetic images. In *CVPR 2024 Workshop SyntaGen: Harnessing Generative Models for Synthetic Visual Datasets*, 2024. 2

[65] Matthias Tangemann, Steffen Schneider, Julius Von Kügelgen, Francesco Locatello, Peter Vincent Gehler, Thomas Brox, Matthias Kuemmerer, Matthias Bethge, and Bernhard Schölkopf. Unsupervised object learning via common fate. In *CLeaR*, 2023. 2

[66] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. StableRep: Synthetic images from text-to-image models make strong visual representation learners. *NeurIPS*, 2023. 2

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 2017. 3

[68] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE T. Image Process.*, 13(4):600–612, 2004. 7

[69] Nick Watters, Loic Matthey, Chris P. Burgess, and Alexander Lerchner. Spatial broadcast decoder: A simple architecture for disentangled representations in VAEs. In *ICLR Workshop on Learning from Limited Labeled Data*, 2019. 3

[70] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. DiffuMask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *ICCV*, pages 1206–1217, 2023. 2, 3, 8

[71] Ziyi Wu, Jingyu Hu, Wuyue Lu, Igor Gilitschenski, and Animesh Garg. SlotDiffusion: Object-centric generative modeling with diffusion models. In *NeurIPS*, 2023. 1, 2, 3, 5

[72] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, and Weidi Xie. Learning open-vocabulary semantic segmentation models from natural language supervision. In *CVPR*, pages 2935–2944, 2023. 8

[73] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, pages 2955–2966, 2023. 2

[74] Yafei Yang and Bo Yang. Promising or elusive? Unsupervised object segmentation from real-world single images. *NeurIPS*, pages 4722–4735, 2022. 2

[75] Andrii Zadaianchuk, Matthaeus Kleindessner, Yi Zhu, Francesco Locatello, and Thomas Brox. Unsupervised semantic segmentation with self-supervised object-centric representations. In *ICLR*, 2023. 2, 8

[76] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[77] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, pages 696–712, 2022. 2, 8

[78] Roland S. Zimmermann, Sjoerd van Steenkiste, Mehdi S.M. Sajjadi, Thomas Kipf, and Klaus Greff. Sensitivity of slot-based object-centric models to their number of slots. *arXiv:2305.18890 [cs.CV]*, 2023. 6