SLOTMASK: MASKED FEATURE RECONSTRUCTION FOR OBJECT-CENTRIC REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans perceive the world as a composition of distinct objects. Inspired by this, object-centric learning represents scenes as sets of object-level vectors called slots. Recent approaches learn slots by reconstructing pretrained visual features with slot-conditioned autoregressive transformers. We introduce SlotMask, which replaces causal reconstruction with a bidirectional transformer trained to predict randomly masked features. Central to our approach is tying masked reconstruction to the object structure discovered by Slot Attention: each masked token is initialized with a mixture of slots weighted by its Slot Attention assignments, and decoder cross-attention is trained to align with these assignments. Together with a randomly sampled masking ratio, these mechanisms make masked reconstruction an effective objective for learning object-centric representations. SlotMask achieves state-of-the-art unsupervised object discovery on PASCAL VOC, MS COCO, and MOVi-C/E, outperforming autoregressive baselines under matched backbones, same training budgets, and with cheaper decoding.

1 INTRODUCTION

Human perception operates on discrete objects: we parse visual scenes into distinct entities, each with its own properties and relationships (Palmer, 1999; Scholl, 2001), an ability that underlies systematic generalization and compositional reasoning (Fodor & Pylyshyn, 1988; Spelke & Kinzler, 2007). Object-centric learning aims to replicate this capacity by representing scenes as sets of feature vectors called *slots* (Greff et al., 2020; Locatello et al., 2020), which have improved sample efficiency in reinforcement learning, systematic generalization, and interpretability (Van Steenkiste et al., 2019; Yoon et al., 2023; Zhang et al., 2025).

Contemporary approaches to object-centric learning often follow a two-stage paradigm (Seitzer et al., 2022; Singh et al., 2022; Kakogeorgiou et al., 2024). First, a pretrained self-supervised vision model, such as DINO (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025), extracts rich features from raw pixels. Second, a Slot Attention (SA) module (Locatello et al., 2020), a differentiable soft K-means clustering with trainable transformations, groups these features into slots and is trained by reconstructing the features with an autoregressive transformer decoder (Vaswani et al., 2017; Radford et al., 2018) conditioned on the slots. This pipeline scales object-centric learning to complex real-world imagery (Seitzer et al., 2022).

Despite the success of causal autoregressive (AR) decoders, masked modeling has emerged as a powerful alternative paradigm in vision, both for self-supervised representation learning (He et al., 2022) and for generative modeling (Chang et al., 2022a; Li et al., 2023; 2024). Instead of predicting tokens one at a time in a fixed order, a masked model predicts randomly masked tokens by conditioning on all visible tokens simultaneously through bidirectional attention. At inference time, it can unmask many tokens in parallel, making decoding cheaper than sequential AR generation. These properties have allowed masked models to match or surpass causal AR models on image generation while remaining strong representation learners.

However, masked feature reconstruction has not been successfully applied to object-centric learning. Recent progress has instead improved DINO feature reconstruction with causal AR decoders (Kakogeorgiou et al., 2024; Kim et al., 2024b; Zhao et al., 2025; Tian et al., 2025; Didolkar et al., 2025), explored slot-conditioned diffusion models (Jiang et al., 2023; Wu et al., 2023; Akan & Yemez, 2025; Singh et al., 2025), or extended object-centric methods to other modalities (Singh

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2022; Manasyan et al., 2025; Gong et al., 2025; Feng et al., 2025). We argue that standard masked reconstruction is poorly matched to slot learning. Every masked token is replaced by the same learnable mask embedding, so it carries no information about the object it belongs to: the decoder must infer this from the visible context, while the Slot Attention assignments, which encode exactly this information, never enter the decoder and are trained only indirectly through the slot contents. Our key observation is that the assignments can serve directly as the decoder input. Initializing each masked token with its assignment-weighted mixture of slots turns the mask token into an object-level hypothesis that the decoder refines, and turns the reconstruction error at every masked position into a direct training signal for the grouping.

In this work, we introduce SlotMask, a slot-conditioned masked feature reconstruction framework for object-centric representation learning. Our contributions are threefold:

- **Slot-aware token initialization**, the key component of the proposed approach. Instead of a shared mask embedding, each masked token is initialized with the mixture of slots given by its Slot Attention assignments, coupling the decoder input to the grouping. It is responsible for the largest part of performance gains, as removing it causes the largest drop in performance (-8.9 mBO_c on COCO), and the additional signal it provides to Slot Attention encoder also improves the learned slot representations.
- **Slot-conditioned masked decoding.** We show that with slot-aware initialization, a bidirectional transformer that reconstructs randomly masked features performs better slots than causal AR decoding, achieving state-of-the-art object-centric performance in both unsupervised object discovery, and set property prediction. At the same time, parallel iterative unmasking requires less decoder forward passes, which makes decoding cheaper.
- **Attention matching and randomized masking.** Our experiments demonstrate the importance of specific design choices, such as the use of randomized masking ratio rather than a fixed one, and aligning decoder cross-attention with the same Slot Attention assignments, so that both the initialization of masked tokens and their updates to the slots follow one object decomposition.

We evaluate SlotMask on real-world (PASCAL VOC, MS COCO) (Everingham et al., 2010; Lin et al., 2014) and synthetic (MOVi-C/E) (Greff et al., 2022) benchmarks. SlotMask achieves state-of-the-art performance in unsupervised object discovery and set property prediction, outperforming AR-based methods under matched backbones and training budgets, and is competitive with a weakly supervised baseline.

2 Related Works

Scene decomposition. Early approaches to unsupervised scene decomposition, such as AIR, NEM, MONet, and IODINE (Eslami et al., 2016; Greff et al., 2017; 2019; Burgess et al., 2019), learned latent-variable generative models that combine iterative inference with mixture-like decoders to separate objects. Slot Attention (SA) (Locatello et al., 2020) subsequently emerged as the dominant approach for extracting slot representations, using soft K-means clustering (Lücke & Forster, 2019) combined with GRU (Cho et al., 2014) and MLP (Rumelhart et al., 1986) updates. The original SA used a mixture decoder that decodes each slot with a Spatial Broadcast decoder (Watters et al., 2019) and combines the outputs with per-slot alpha masks; this pixel-level objective works on synthetic data but struggles on complex, texture-rich images.

Autoregressive object-centric models. SLATE (Singh et al., 2021) addressed limitations of pixel-level autoencoders by using an autoregressive transformer conditioned on slots to reconstruct dVAE-encoded image tokens (Van Den Oord et al., 2017), though it still struggled with real-world images. DINOSAUR (Seitzer et al., 2022) overcame this challenge by using pretrained self-supervised vision transformer DINO (Caron et al., 2021) features as the reconstruction target, enabling learning of meaningful object-centric representations on complex real-world datasets. SPOT (Kakogeorgiou et al., 2024) further improved this approach through iterative self-training and prediction order permutation.

Compositional generation with diffusion models. A complementary line of work uses slot-conditioned diffusion models for compositional generation. SlotDiffusion (Wu et al., 2023) trains from scratch, Stable-LSD (Jiang et al., 2023) leverages a pretrained model, and SlotAdapt (Akan &

Yemez, 2025) fine-tunes a pretrained model through adapter layers. Moving beyond unsupervised learning, GLASS (Singh et al., 2025) uses ground-truth labels to generate captions that guide Slot Attention training.

Masked learning. Masking portions of the input and training a model to reconstruct the masked content from the visible context has emerged as a powerful self-supervised paradigm for vision, pioneered by Masked Autoencoders (MAE) (He et al., 2022). MaskGIT (Chang et al., 2022a) extended it to generative modeling with bidirectional attention and parallel prediction of multiple tokens, outperforming autoregressive baselines, and MAGE (Li et al., 2023) showed that it excels at both generation and representation learning. MAR (Li et al., 2024) further unified masked and autoregressive generation and extended it to continuous tokens via a diffusion loss. To our knowledge, MOGENT (Nakano et al., 2025) is the only masked-modeling approach to object-centric learning to date, yet it falls short of state-of-the-art AR and diffusion-based alternatives. SlotMask borrows from masked generative models only the bidirectional decoder, the randomly sampled masking ratio, and parallel iterative unmasking, and keeps the simple ℓ_2 feature reconstruction objective of DINOSAUR without a diffusion loss or discrete tokens. Its key difference from all these methods is how masked tokens are formed: instead of a shared learnable mask token, each masked token is initialized from the slots according to the Slot Attention assignments.

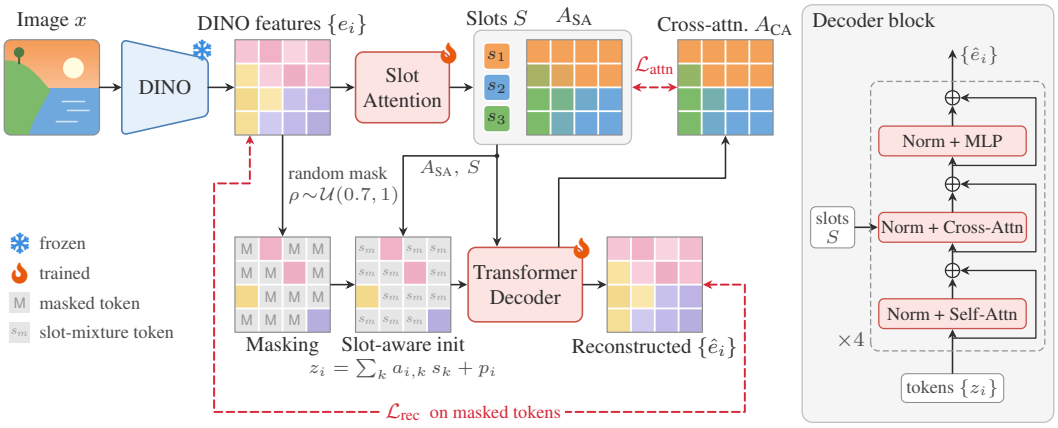

Figure 1: Overview of SlotMask. Given an input image, a frozen DINOv3 ViT encoder extracts patch tokens $\{e_i\}$. Slot Attention groups these tokens into K slot representations $\{s_k\}$. We then randomly mask DINO features and perform slot-aware token initialization by replacing masked tokens with slot mixtures $\sum_k a_{i,k} s_k$ (plus positional embeddings) and feed the resulting sequence to a slot-conditioned transformer decoder that reconstructs the masked DINO tokens, trained with an ℓ_2 reconstruction loss $\mathcal{L}_{\text{rec}}$ on masked tokens, and an attention-matching loss $\mathcal{L}_{\text{attn}}$ aligning Slot Attention assignments A_{SA} with decoder cross-attention scores A_{CA}.

3 Method

SlotMask learns object-centric representations in two stages. First, we perform *Slot Encoding* (Section 3.1): a frozen pretrained encoder extracts visual tokens, and Slot Attention groups them into K object-centric slots. Second, we perform *Slot-Conditioned Masked Decoding* (Section 3.2): we sample a random token mask, replace the masked tokens with slot-aware initializations, and use a slot-conditioned masked decoder to reconstruct the masked visual tokens. Our main technical contribution is *slot-aware token initialization*, which feeds the Slot Attention assignments into the decoder through the masked tokens. See Fig. 1 for a visual overview of the proposed approach.

3.1 SLOT ENCODING

Pretrained visual tokens. We encode input image $x \in \mathbb{R}^{H\times W\times 3}$ using a frozen DINOv3 ViT encoder f_{DINO}, obtaining a grid of N patch tokens $E = \{e_i\}_{i=1}^{N} = f_{\text{DINO}}(x) \in \mathbb{R}^{N\times D}$, where $N = HW/P^2$ for patch size P and D is the token dimensionality.

Slot Attention. Slot Attention maps E to a set of K slots $S \in \mathbb{R}^{K\times D_s}$ and outputs a soft assignment matrix $A_{\text{SA}} \in [0,1]^{N\times K}$ indicating which slot explains each token. Slots are initialized as $s_k^{(0)} \sim \mathcal{N}(\mu, \text{diag}(\sigma^2))$, with trainable $\mu, \sigma \in \mathbb{R}^{D_s}$. For T refinement iterations, Slot Attention alternates between computing assignments and updating slots. Using projections $q_k^{(t)} = W_q s_k^{(t)}, \quad k_i = W_k e_i, \quad v_i = W_v e_i$, the slots are updated as

$$a_{i,k}^{(t)} = \frac{\exp\left(\langle k_i, q_k^{(t)}\rangle/\sqrt{D_s}\right)}{\sum_{k'=1}^{K}\exp\left(\langle k_i, q_{k'}^{(t)}\rangle/\sqrt{D_s}\right)}, \quad u_k^{(t)} = \sum_{i=1}^{N}\frac{a_{i,k}^{(t)}}{\sum_{i'=1}^{N} a_{i',k}^{(t)} + \epsilon} v_i, \tag{1}$$

weighted average values $u_k^{(t)}$ are updated with GRU and MLP modules:

$$s_k^{(t+1)} = \text{MLP}\Big(\text{GRU}\Big(s_k^{(t)},\, u_k^{(t)}\Big)\Big). \tag{2}$$

We use the final slots $S = \{s_k^{(T)}\}_{k=1}^{K}$ and final assignments $A_{\text{SA}} = \{a_{i,k}^{(T)}\}_{i,k}$. Following prior work, we adopt standard Slot Attention stabilization techniques for training (Chang et al., 2022b).

3.2 SLOT-CONDITIONED MASKED DECODER

Masked decoding. For each training batch, we sample a masking ratio $\rho \sim \mathcal{U}(\rho_{\text{min}}, 1)$ with $\rho_{\text{min}} = 0.7$ and generate a binary mask $m \in \{0,1\}^N$, where $m_i = 1$ indicates token i is masked. Let $\mathcal{M} = \{i : m_i = 1\}$ and $\mathcal{V} = \{j : m_j = 0\}$ denote masked and visible token indices. The decoder learns to reconstruct masked tokens conditioned on visible ones and on the slots:

$$E_{\mathcal{M}} \sim p\left(E_{\mathcal{M}} \,|\, E_{\mathcal{V}}, S\right). \tag{3}$$

Sampling ρ rather than fixing it, as in MAE, matters for slot learning: a fixed ratio lets the decoder rely on a predictable amount of visible context instead of the slots, whereas a sampled ratio exposes the decoder to every conditioning regime encountered during iterative decoding, from almost fully masked to mostly visible. Fixed ratios perform substantially worse, see Section A.1.

Slot-aware masked token initialization. A crucial difference from standard masked modeling, where every masked token is replaced by the same learnable mask embedding plus a positional embedding, is that we perform *mask filling*: each masked token is initialized with a *slot mixture* defined by the Slot Attention assignments:

$$z_i = \begin{cases} e_i, & i \in \mathcal{V}, \\ \sum_{k=1}^{K} a_{i,k}\, s_k \;+\; p_i, & i \in \mathcal{M}, \end{cases} \tag{4}$$

where p_i is the learnable positional embedding at location i. Compared with a shared mask embedding, this has two effects. First, each masked token starts from an object-level estimate, the slots of the objects it was assigned to, so the decoder refines a hypothesis about location i instead of inferring its object membership from the visible context. Second, because z_i depends on $a_{i,k}$, the reconstruction error at every masked token backpropagates into its assignment, so the loss supervises *where* each slot is, not only *what* it contains; a decoder that reads the slots only through cross-attention reaches the assignments only indirectly, through the slot contents. Slot-aware initialization thus moves the per-location slot mixture of the original Slot Attention decoder (Locatello et al., 2020) from the output to the input of a transformer decoder, with weights given by Slot Attention instead of predicted by the decoder, keeping an explicit binding between slots and locations inside an expressive decoder.

Slot-conditioned transformer decoder. Our decoder is a bidirectional transformer with L layers. Each layer ℓ applies multi-head self-attention over token embeddings $Z^{(\ell)} \in \mathbb{R}^{N\times D}$, followed by

multi-head cross-attention from tokens (queries) to slots (keys/values), and a final MLP. For head h, queries are computed from tokens as $q^{(\ell,h)} = Z^{(\ell)} w_q^{(\ell,h)}$, while keys and values are computed from slots $S \in \mathbb{R}^{K \times D_s}$ as $k^{(\ell,h)} = S w_k^{(\ell,h)}$ and $v^{(\ell,h)} = S w_v^{(\ell,h)}$, where $w_q^{(\ell,h)}, w_k^{(\ell,h)}, w_v^{(\ell,h)}$ are learned projection matrices and d_h is the per-head dimensionality. The token-to-slot cross-attention weights are then

$$A_{\text{CA}}^{(\ell,h)} = \text{softmax}\left(\frac{q^{(\ell,h)} k^{(\ell,h)\top}}{\sqrt{d_h}}\right) \in [0,1]^{N \times K}, \tag{5}$$

where the softmax is taken over the slot dimension. We use these cross-attention scores as a decoder segmentation mask by averaging across all heads and layers.

Reconstruction loss. We use an ℓ_2 loss on masked tokens (as in prior DINO-feature reconstruction methods (Seitzer et al., 2022; Kakogeorgiou et al., 2024)):

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\hat{e}_i - e_i\|_2^2 . \tag{6}$$

Attention matching loss. To also tie how the decoder reads the slots to the assignments that initialize its input, we add an attention matching term (inspired by SlotAdapt (Akan & Yemez, 2025)). We then minimize cross-entropy between A_{SA} and A_{CA}:

$$\mathcal{L}_{\text{attn}} = -\sum \text{sg}(A_{\text{SA}}) \log A_{\text{CA}} - \sum \text{sg}(A_{\text{CA}}) \log A_{\text{SA}}, \tag{7}$$

where sg() denotes stop gradient. The total objective is

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{attn}}, \tag{8}$$

with $\lambda = 0.1$. We linearly warm up λ from 0 to 0.1 over the first 40k training steps.

Iterative masked decoding and cost. As in prior work, decoder cross-attention scores A_{CA} give marginally better object masks than SA assignments (Seitzer et al., 2022; Kakogeorgiou et al., 2024), and iterative unmasking improves them slightly further, see Section A.3. We follow the "predict-the-next-set" procedure of masked generative models (Chang et al., 2022a; Li et al., 2024): starting from all tokens masked, we iteratively unmask tokens, decreasing the masking ratio from 1 to 0 over 64 steps with a cosine schedule. This remains cheaper than causal AR decoding, see Section A.4.

4 Experiments

4.1 Setup

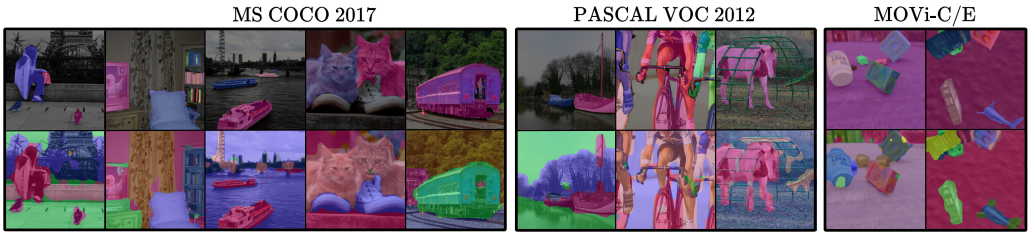

Figure 2: Qualitative examples from three datasets. The first row shows ground-truth instance masks, the second row visualizes the slot masks by our method (SlotMask).

Datasets. We use both real-world and synthetic datasets. For real-world data, we use MS COCO 2017 (Lin et al., 2014) with its official splits (118 287 training and 5 000 validation images) and PASCAL VOC 2012 (Everingham et al., 2010), training on the *trainaug* split (10 582 images including SBD (Hariharan et al., 2011)) and validating on the original 1 449 validation images. VOC is the easier benchmark, as it contains fewer objects per image. Following prior work (Seitzer et al., 2022; Kakogeorgiou et al., 2024), we also use frames from the synthetic MOVi-C and MOVi-E video datasets (Greff et al., 2022), which contain 3–10 and 11–23 realistic objects per scene, respectively.

Table 1: Quantitative comparison of object discovery performance across datasets. Reported values are mean ± std for 5 random seeds when available. Best results are in **bold** and second-best results are underlined. **Note: GLASS is the only baseline that is not fully unsupervised, it uses ground-truth labels to generate captions that guide Slot Attention training, providing a weak supervision signal.*

	COCO		VOC		MOVi-E		MOVi-C	
Model	mBO_i	mBO_c	mBO_i	mBO_c	$mIoU_i$	mBO_i	$mIoU_i$	mBO_i
MLP decoders								
SA	17.2	19.2	24.6	24.9	–	24.0 ± 1.2	–	26.2 ± 1.0
DINOSAUR (MLP)	27.7 ± 0.2	30.9 ± 0.2	39.5 ± 0.1	40.9 ± 0.1	–	35.5 ± 0.2	–	39.1 ± 0.2
AR decoders								
SLATE	29.1	33.6	35.9	41.5	28.6 ± 1.7	30.2 ± 1.7	37.8 ± 0.7	39.4 ± 0.8
DINOSAUR (AR)	32.3 ± 0.4	38.8 ± 0.4	44.0 ± 1.9	51.2 ± 1.9	–	–	–	42.4
$DINOSAUR_{v3}$	34.7 ± 0.5	42.3 ± 0.4	46.0 ± 2.1	53.3 ± 2.5	37.6	39.0	44.7	44.0
SPOT	35.0 ± 0.1	44.7 ± 0.3	48.3 ± 0.4	55.6 ± 0.4	39.3 ± 1.2	40.1 ± 1.2	46.7 ± 1.3	47.3 ± 1.2
$SPOT_{v3}$	35.1 ± 0.2	45.9 ± 0.2	48.9 ± 0.7	57.5 ± 0.9	40.0	41.1	47.0	48.0
Diffusion decoders								
SlotAdapt	35.1	39.2	51.5	51.9	41.85	43.38	–	–
GLASS*	**40.8**	48.5	**58.9**	62.2	–	–	–	–
SlotMask **(ours)**	36.1 ± 0.2	**48.8** ± 0.3	58.6 ± 1.5	**69.8** ± 2.0	**42.0** ± 0.7	**45.5** ± 1.4	**47.8** ± 0.9	**49.5** ± 0.9

Evaluation metrics. We compare slot masks with ground-truth instance and semantic segmentations using Mean Best Overlap (mBO_i and mBO_c for instance and semantic masks) and mean Intersection over Union (mIoU). mBO matches each ground-truth mask to the best-overlapping slot, while mIoU uses Hungarian matching (Kuhn, 1955). For COCO and VOC, we report mBO_i and mBO_c, the most widely reported metrics, which enable direct comparison to many baselines; the MOVi datasets focus on instance segmentation, so we report $mIoU_i$ and mBO_i. We do not use Foreground Adjusted Rand Index (FG-ARI) (Singh et al., 2021; Seitzer et al., 2022; Singh et al., 2025), since it often does not reflect mask or representation quality and can reward over-segmentation (Kakogeorgiou et al., 2024).

Training setup. Unless stated otherwise, we use a frozen DINOv3 (Siméoni et al., 2025) encoder of size B to extract patch-level features. Slot Attention uses $K = 7$ and $K = 6$ slots for COCO and VOC, respectively, following the most commonly used settings in prior work. For MOVi-C/E, we use $K = 11/24$, corresponding to the maximum number of objects per scene plus one slot for background. Slot dimensionality 512 and numbe of SA iterations is 5, we use the fixed-point optimization trick from (Chang et al., 2022b) for more stable training. The slot-conditioned masked decoder is a 4-layer transformer with hidden size 768 (matching the DINOv3-B token dimensionality). We train our models with Adam using learning rate 3×10^{-4} for COCO and VOC datasets, and 1.5×10^{-4} for MOVi-C/E. We linearly warm up the learning rate from 0 for the first 10k steps, and then decay it with a cosine schedule to a minimum learning rate of 10^{-6}. We train for 150k iterations on COCO and VOC and for 100k iterations on MOVi-C/E (compared to the more common 300k–500k iterations used by baselines (Seitzer et al., 2022; Singh et al., 2025; Kakogeorgiou et al., 2024)), with batch size 64 for all experiments. We use images of size 256×256 by default. We upsample slot masks from the 16×16 patch grid to the image resolution before computing the metrics with bilinear interpolation. All experiments were run on a single NVIDIA RTX A6000. A complete code implementation with training/validation scripts and detailed config files are provided in the code appendix.

Baselines. We group baselines by decoder type. Slot Attention (SA) (Locatello et al., 2020) and DINOSAUR (MLP) (Seitzer et al., 2022) use MLP mixture decoders, reconstructing pixels and frozen DINO features, respectively. SLATE (Singh et al., 2021), DINOSAUR (AR) (Seitzer et al., 2022), and SPOT (Kakogeorgiou et al., 2024) use slot-conditioned causal AR transformer decoders; SLATE reconstructs discrete dVAE tokens, while DINOSAUR and SPOT reconstruct DINO features, with SPOT adding prediction-order permutations and self-training. For a fully controlled comparison with AR baselines, we retrain DINOSAUR and SPOT with the same frozen DINOv3-B encoder used by SlotMask, matching slot counts, training budgets, optimization schedules, and mask-extraction

procedures, and report mean ± std over 5 seeds; we refer to these variants as DINOSAUR$_{v3}$ and SPOT$_{v3}$. SlotAdapt (Akan & Yemez, 2025) is the strongest unsupervised diffusion-based baseline. We additionally compare to the weakly supervised diffusion-based GLASS (Singh et al., 2025), which uses ground-truth labels to generate captions whose attention maps serve as pseudo-masks for training; SlotMask and all other baselines are fully unsupervised. For the remaining baselines we report numbers from the original papers ("–" where a metric was not reported), with mean ± std over 5 seeds where available. For qualitative comparison (see Fig. 3) and property prediction training, we use the official checkpoints for SPOT and GLASS.

4.2 Results

Main comparison. Table 1 presents instance and semantic object discovery results. Our unsupervised SlotMask demonstrates state-of-the-art performance on both instance and semantic object discovery, outperforming all unsupervised baselines across all four datasets. It also remains competitive with the weakly supervised GLASS, even outperforming it on mBO$_c$, while using fewer training iterations (100–150k vs. 300–500k). Figure 2 shows qualitative examples of masks produced by our approach alongside the ground-truth masks across datasets. See Figure 3 for a qualitative comparison of COCO slot masks produced by SlotMask, GLASS, and SPOT.



Figure 3: Qualitative comparison of slot masks produced by SlotMask, SPOT, and GLASS on the MS COCO 2017 dataset.

Property prediction. Table 2 reports results on the instance-level property prediction task. We freeze the pretrained encoder and Slot Attention module to extract slots, and then train a 3-layer MLP on top of these slots. The MLP takes a slot as input and predicts the object class label as well as the 2D coordinates of the bounding-box center, with each coordinate normalized to $[0, 1]$. We match predicted slots to the ground-truth set of object properties using Hungarian matching (Kuhn, 1955).

Table 2: Object property prediction on real-world datasets. We predict the class label (Acc, higher is better) and the 2D bounding-box center position with coordinates $\in [0, 1]$ (MSE $\times 10^{-2}$, lower is better).

Model	COCO		VOC	
	Acc (%) ↑	MSE ↓	Acc (%) ↑	MSE ↓
GLASS	20.8	5.9	58.1	3.7
SPOT	71.1 ± 1.4	2.7 ± 0.2	80.6 ± 1.3	2.5 ± 0.3
SPOT$_{v3}$	74.3 ± 1.7	**2.5** ± 0.3	83.1	**2.4** ± 0.2
SlotMask	**80.8** ± 1.5	**2.5** ± 0.2	**87.2** ± 2.3	**2.4** ± 0.2

Table 3: Ablation on MS COCO: influence of architectural design choices on performance.

Backbone	Slot init	Attn	Dec	Steps	λ	ρ_{min}	mBO$_i$ ↑	mBO$_c$ ↑
v3-S	✓	full	masked	64	0.1	0.7	34.0	44.6
v3-B	✓	full	masked	64	0.1	0.7	**36.1**	**48.8**
v3-L	✓	full	masked	64	0.1	0.7	35.7	48.3
v3-B	×	full	masked	64	0.1	0.7	34.7	39.9
v3-B	✓	causal	masked	64	0.1	0.7	35.5	47.6
v3-B	✓	full	AR	256	0.1	0.7	35.0	46.4
v3-B	✓	full	masked	1	0.1	0.7	35.2	47.7
v3-B	✓	full	masked	64	0.05	0.7	35.1	46.8
v3-B	✓	full	masked	64	0.2	0.7	35.2	47.2
v3-B	✓	full	masked	64	0.5	0.7	33.6	43.7
v3-B	✓	full	masked	64	0.1	0.6	35.8	47.5
v3-B	✓	full	masked	64	0.1	0.8	35.0	47.3

We report top-1 accuracy for label prediction (Acc) and mean squared error for coordinate prediction ($\text{MSE} \times 10^{-2}$). The image-generation-focused GLASS performs substantially worse than the unsupervised representation learning methods on both tasks. On COCO, GLASS obtains 20.8% accuracy and 5.9 MSE, compared with $71.1\%/2.7$, $74.3\%/2.5$, and $80.8\%/2.5$ for SPOT, SPOT$_{\text{v3}}$, and SlotMask, respectively. SlotMask achieves the highest accuracy while matching the best coordinate MSE with SPOT$_{\text{v3}}$.

Design ablation Table 3 studies how architectural and objective choices affect SlotMask on MS COCO. We ablate (i) the size of the DINO backbone, (ii) slot-aware token initialization, (iii) causal vs. bidirectional attention in the decoder, (iv) masked vs. causal AR decoding, (v) single-pass decoding, (vi) the weight λ of the attention-matching loss, and (vii) the minimum masking ratio ρ_{min}. Slot-aware initialization is the most important component: removing it lowers mBO$_c$ by 8.9 points, far more than replacing masked decoding with causal AR decoding (2.4 points), so coupling the decoder input to the Slot Attention assignments is the main source of SlotMask's gains. Masked decoding outperforms causal AR decoding even with a single decoding pass, and bidirectional attention outperforms causal attention. SlotMask is relatively robust to moderate changes in λ and ρ_{min}. Varying the DINO backbone yields non-uniform effects: performance drops with the smaller DINOv3-S, but scaling up to DINOv3-L does not help and is less stable during training (see Section 5).

Cross-dataset performance Table 4 compares target-trained models with COCO-trained models evaluated zero-shot on VOC and MOVi-E. On VOC, SlotMask transfers almost without loss (58.8 vs. 58.6 mBO$_i$, 69.1 vs. 69.8 mBO$_c$), and SPOT and GLASS show similarly small drops. On MOVi-E, all COCO-trained models degrade substantially due to the shift from real to synthetic scenes; SlotMask and SPOT show comparable relative drops, whereas GLASS is markedly worse in absolute terms, suggesting that self-supervised methods generalize more robustly than the weakly supervised GLASS.

Unsupervised semantic segmentation on COCO-Stuff. COCO-Stuff (Caesar et al., 2018) extends MS COCO with per-pixel labels for both "thing" and "stuff" classes (e.g., sky, road, grass). We use it to test whether our object-centric representations translate into strong pixel-level semantic segmentation. Since slot masks are produced on the 16×16 patch grid and bilinearly upsampled to 256×256, boundaries can be imprecise; we therefore also report a variant refined with a Conditional Random Field (Krähenbühl & Koltun, 2011) (CRF), a training-free post-processing step widely used in unsupervised segmentation (Wang et al., 2023b;a; Kim et al., 2024a; Niu et al., 2024; Curie & da Costa, 2025). Table 5 reports semantic mean IoU (mIoU$_c$) and Pixel Accuracy against strong non-slot-based unsupervised baselines: STEGO (Hamilton et al., 2022), EAGLE (Kim et al., 2024a), U2Seg (Niu et al., 2024), and CLASP (Curie & da Costa, 2025). SlotMask outperforms all baselines on both metrics, and CRF refinement boosts performance further, indicating the value of boundary refinement when transferring patch-level slot assignments to pixel-level segmentations.

Table 4: Cross-dataset performance. We compare models trained on the target datasets (VOC and MOVi-E) against models trained only on COCO and evaluated on VOC and MOVi-E.

		VOC		MOVi-E	
Model	**Train set**	$\text{mBO}_i \uparrow$	$\text{mBO}_c \uparrow$	$\text{mIoU}_i \uparrow$	$\text{mBO}_i \uparrow$
SlotMask	target	$58.6_{\pm 1.5}$	$69.8_{\pm 2.0}$	$42.0_{\pm 0.7}$	$45.5_{\pm 1.4}$
SlotMask	COCO	$58.8_{\pm 1.7}$	$69.1_{\pm 2.2}$	$33.2_{\pm 0.5}$	$36.8_{\pm 0.5}$
SPOT	target	$48.3_{\pm 0.4}$	$55.6_{\pm 0.4}$	$39.3_{\pm 1.2}$	$40.1_{\pm 1.2}$
SPOT	COCO	$48.0_{\pm 0.6}$	$54.9_{\pm 0.7}$	$31.2_{\pm 0.8}$	$32.4_{\pm 0.9}$
GLASS	target	58.9	62.2	–	–
GLASS	COCO	57.7	61.8	24.8	27.9

Table 5: Comparison with models for unsupervised semantic segmentation on COCO-Stuff.

Metric	**STEGO**	**EAGLE**	**U2Seg**	**SlotMask$_{\text{RN50}}$**	**CLASP**	**SlotMask$_{\text{v2-S}}$**	**SlotMask**	**SlotMask$_{\text{CRF}}$**
$\text{mIoU}_c \uparrow$	28.2	27.2	30.2	33.4	36.1	36.3	<u>36.6</u>	**38.5**
PixelAcc ↑	56.9	59.4	63.9	66.2	64.4	68.3	<u>70.7</u>	**72.4**

Additional analyses. We summarize further experiments reported in the appendix. We find randomized masking ratio to be essential: every fixed ratio from 0.75 to 1.0 underperforms $\rho \sim \mathcal{U}(0.7, 1)$ (Section A.1). We also ablate the pretrained encoder: with either DINOv1 or DINOv3 features, SlotMask outperforms the corresponding baseline variants (Section A.2). Raw Slot Attention masks of SlotMask only slightly lag behind the decoder masks and outperform the decoder masks of all the baselines (Section A.3). Unlike SPOT$_{\text{v3}}$, SlotMask keeps improving as the number of slots increases (Section A.5). SlotMask also uses fewer active slots with lower assignment entropy than SPOT$_{\text{v3}}$ (Section A.6). Finally, we show the importance of slot-aware initialization with additional property prediction experiments (Section A.7).

5 Limitations

SlotMask shares several limitations with existing object-centric methods. First, slot-attention models require a fixed slot budget K, which is dataset-dependent and can hurt performance when the number of objects varies across images. Although we follow standard choices, such as $K = 7$ for COCO and larger K for MOVi-E, adaptive slot allocation remains an open problem (Fan et al., 2024; Liu et al., 2025; Ouyang et al., 2026). Second, SlotMask relies on a strong frozen pretrained encoder, and stronger encoders do not automatically improve object discovery: scaling from DINOv3-B to DINOv3-L did not improve performance and sometimes introduced training instabilities. A similar effect has been reported for generative modeling on pretrained DINO features (Chefer et al., 2026); we hypothesize that larger encoders produce higher-dimensional, more anisotropic feature spaces that are harder to reconstruct under an ℓ_2 objective and to compress into a single slot vector.

6 Conclusion

We introduced SlotMask, a masked feature reconstruction framework for object-centric learning whose central idea, slot-aware token initialization, uses the Slot Attention assignments as the decoder input for masked tokens, so that masked reconstruction both starts from and directly trains the object decomposition. Across real-world and synthetic benchmarks, it achieves state-of-the-art object discovery, improves downstream property prediction, and decodes faster than autoregressive baselines. These results show that masked feature reconstruction, once tied to the object structure found by Slot Attention, is a strong alternative to causal autoregression for learning object-centric representations.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the implementation of the proposed method and the experimental pipeline, as well as for language polishing and grammar correction of the manuscript. All AI-assisted code was carefully reviewed and tested by the authors for correctness and faithfulness to the method described in this paper, and all AI-assisted edits to the text were reviewed by the authors. We have not used generative AI tools for research ideation or the generation of the scientific contributions of this work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The architecture and training objective are described in Section 3, and datasets, evaluation protocols, and hyperparameters are given in the experimental setup (Section 4.1). We build on publicly available pretrained DINO encoders and standard benchmarks. We include the source code and the configuration files for all experiments in the supplementary material, allowing full reproduction of the reported results.

REFERENCES

Adil Kaan Akan and Yucel Yemez. Slot-guided adaptation of pre-trained diffusion models for object-centric learning and compositional generation. *arXiv preprint arXiv:2501.15878*, 2025.

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.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

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

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022a.

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

Hila Chefer, Patrick Esser, Dominik Lorenz, Dustin Podell, Vikash Raja, Vinh Tong, Antonio Torralba, and Robin Rombach. Self-supervised flow matching for scalable multi-modal synthesis. *arXiv preprint arXiv:2603.06507*, 2026.

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. *arXiv preprint arXiv:1406.1078*, 2014.

Max Curie and Paulo da Costa. Clasp: Adaptive spectral clustering for unsupervised per-image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5982–5991, 2025.

Aniket Didolkar, Andrii Zadaianchuk, Rabiul Awal, Maximilian Seitzer, Efstratios Gavves, and Aishwarya Agrawal. Ctrl-o: language-controllable object-centric visual representation learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29523–29533, 2025.

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

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*, 88(2): 303–338, 2010.

Ke Fan, Zechen Bai, Tianjun Xiao, Tong He, Max Horn, Yanwei Fu, Francesco Locatello, and Zheng Zhang. Adaptive slot attention: Object discovery with dynamic slot number. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23062–23071, 2024.

Fan Feng, Phillip Lippe, and Sara Magliacane. Learning interactive world model for object-centric reinforcement learning. *arXiv preprint arXiv:2511.02225*, 2025.

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

Xinrui Gong, Oliver Hahn, Christoph Reich, Krishnakant Singh, Simone Schaub-Meyer, Daniel Cremers, and Stefan Roth. Motion-refined dinosaur for unsupervised multi-object discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 220–230, 2025.

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

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

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

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3749–3761, 2022.

Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*, 2022.

Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pp. 991–998. IEEE, 2011.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. *arXiv preprint arXiv:2303.10834*, 2023.

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

Chanyoung Kim, Woojung Han, Dayun Ju, and Seong Jae Hwang. Eagle: Eigen aggregation learning for object-centric unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3523–3533, 2024a.

Dongwon Kim, Seoyeon Kim, and Suha Kwak. Bootstrapping top-down information for self-modulating slot attention. *Advances in Neural Information Processing Systems*, 37:103751–103773, 2024b.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2142–2152, 2023.

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

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 *European conference on computer vision*, pp. 740–755. Springer, 2014.

Hongjia Liu, Rongzhen Zhao, Haohan Chen, and Joni Pajarinen. Metaslot: Break through the fixed number of slots in object-centric learning. *arXiv preprint arXiv:2505.20772*, 2025.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

Jörg Lücke and Dennis Forster. k-means as a variational em approximation of gaussian mixture models. *Pattern Recognition Letters*, 125:349–356, 2019.

Anna Manasyan, Maximilian Seitzer, Filip Radovic, Georg Martius, and Andrii Zadaianchuk. Temporally consistent object-centric learning by contrasting slots. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5401–5411, 2025.

Akihiro Nakano, Masahiro Suzuki, and Yutaka Matsuo. Efficient object-centric representation learning using masked generative modeling. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=t9KvOYPeL3.

Dantong Niu, Xudong Wang, Xinyang Han, Long Lian, Roei Herzig, and Trevor Darrell. Unsupervised universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22744–22754, 2024.

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 preprint arXiv:2304.07193*, 2023.

Tianran Ouyang, Xingping Dong, Jing Zhang, Mang Ye, Jun Chen, and Bo Du. Qasa: Quality-guided k-adaptive slot attention for unsupervised object-centric learning. *arXiv preprint arXiv:2601.12936*, 2026.

Stephen E Palmer. *Vision science: Photons to phenomenology*. MIT press, 1999.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Brian J Scholl. Objects and attention: The state of the art. *Cognition*, 80(1-2):1–46, 2001.

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. *arXiv preprint arXiv:2209.14860*, 2022.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Gautam Singh, Fei Deng, and Sungjin Ahn. Illiterate dall-e learns to compose. *arXiv preprint arXiv:2110.11405*, 2021.

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

Krishnakant Singh, Simone Schaub-Meyer, and Stefan Roth. Glass: Guided latent slot diffusion for object-centric learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28673–28683, 2025.

Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.

Pinzhuo Tian, Shengjie Yang, Hang Yu, and Alex Kot. Pay attention to the foreground in object-centric learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30281–30290, 2025.

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

Sjoerd Van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? *Advances in neural information processing systems*, 32, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3124–3134, 2023a.

Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *IEEE transactions on pattern analysis and machine intelligence*, 45(12):15790–15801, 2023b.

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.

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

Jaesik Yoon, Yi-Fu Wu, Heechul Bae, and Sungjin Ahn. An investigation into pre-training object-centric representations for reinforcement learning. *arXiv preprint arXiv:2302.04419*, 2023.

Weipu Zhang, Adam Jelley, Trevor McInroe, and Amos Storkey. Objects matter: object-centric world models improve reinforcement learning in visually complex environments. *arXiv preprint arXiv:2501.16443*, 2025.

Rongzhen Zhao, Yi Zhao, Juho Kannala, and Joni Pajarinen. Slot attention with re-initialization and self-distillation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4185–4192, 2025.

A Additional Experiments

All experiments in this appendix use the default SlotMask configuration on MS COCO ($K = 7$, DINOv3-B) unless stated otherwise, and report decoder masks after 64 decoding steps.

A.1 MASKING-RATIO STRATEGY

Table 6 compares fixed masking ratios, as used in MAE (He et al., 2022), against sampling the ratio per batch. The standard MAE ratio of 0.75 is far from effective for slot learning: with this much visible context, the decoder can reconstruct the input without relying on the slots. Increasing the fixed ratio helps, but even the best fixed ratio (0.90) remains well below $\rho \sim \mathcal{U}(0.7, 1)$. Sampling the ratio exposes the decoder to the full range of conditioning regimes encountered during iterative decoding, from almost fully masked to more visible, and encourages it to make consistent use of the slots.

Table 6: Fixed vs. sampled masking ratios on MS COCO.

Masking ratio ρ	$\text{mBO}_i \uparrow$	$\text{mBO}_c \uparrow$
fixed 0.75	25.9	34.1
fixed 0.80	27.5	36.0
fixed 0.85	30.4	39.4
fixed 0.90	34.6	45.8
fixed 0.95	34.2	45.1
fixed 1.00	33.2	42.5
$\mathcal{U}(0.6, 1)$	35.8	47.5
$\mathcal{U}(0.7, 1)$ (default)	**36.1**	**48.8**
$\mathcal{U}(0.8, 1)$	35.0	47.3

A.2 CONTROLLING FOR THE PRETRAINED BACKBONE

To disentangle the contribution of the pretrained features from that of our method, Table 7 reports K-means clustering of frozen DINOv1 and DINOv3 patch features (K-Means$_{\text{v1}}$ and K-Means$_{\text{v3}}$), the AR baselines, and SlotMask trained on either backbone. Within each backbone, SlotMask outperforms K-means and both AR baselines, confirming that the gains are not attributable to DINOv3 alone. Notably, SlotMask$_{\text{v1}}$ already matches or exceeds SPOT$_{\text{v3}}$ on most metrics.

Table 7: Object discovery with matched backbones. All models within a block use the same frozen encoder.

		COCO		**VOC**	
Backbone	**Model**	$\text{mBO}_i \uparrow$	$\text{mBO}_c \uparrow$	$\text{mBO}_i \uparrow$	$\text{mBO}_c \uparrow$
DINOv1	K-Means	25.5	32.2	41.2	45.6
DINOv1	DINOSAUR	32.3	38.8	44.0	51.2
DINOv1	SPOT	35.0	44.7	48.3	55.6
DINOv1	SlotMask	**35.2**	**45.8**	**50.2**	**58.4**
DINOv3	K-Means	25.9	34.7	42.7	49.5
DINOv3	DINOSAUR	34.7	42.3	46.0	53.3
DINOv3	SPOT	35.1	45.9	48.9	57.5
DINOv3	SlotMask	**36.1**	**48.8**	**58.6**	**69.8**

A.3 SLOT ATTENTION MASKS VS. DECODER MASKS

Following common practice (Seitzer et al., 2022; Kakogeorgiou et al., 2024), the main results use decoder cross-attention masks. Table 8 compares them with the raw Slot Attention assignment masks with decoder masks. Decoder masks are slightly more precise, but the Slot Attention masks alone already outperform the decoder masks of almost all baselines' decoder masks in Table 1.

A.4 DECODING COST

Because the SlotMask decoder is bidirectional, one forward pass predicts all masked tokens at once, and iterative decoding needs only as many passes as unmasking steps. Causal AR decoders, as in DINOSAUR and SPOT and the AR variant of our decoder in Table 3, instead generate the

Table 8: Mask source comparison for SlotMask.

	COCO		VOC		MOVi-E		MOVi-C	
Mask source	mBO_i	mBO_c	mBO_i	mBO_c	mIoU_i	mBO_i	mIoU_i	mBO_i
Slot Attention	34.9	47.2	56.5	68.9	41.9	45.2	47.5	48.7
Decoder	**36.1**	**48.8**	**58.6**	**69.8**	**42.0**	**45.5**	**47.8**	**49.5**

$16 \times 16 = 256$ tokens one at a time and need 256 sequential passes. Table 9 reports the inference cost of each decoding mode together with its mask quality on MS COCO. A single decoding pass costs $\approx$20 GFLOPs, and 64-step decoding $\approx$500 GFLOPs. This is well below $64\times$ the single-pass cost because each step processes only the tokens decoded so far and those being predicted, so early steps operate on short sequences. Sequential AR decoding costs $\approx$5,000 GFLOPs. Even the most expensive SlotMask setting is thus cheaper than AR decoding in both FLOPs and wall-clock time, while producing better masks. Single-step decoding is cheaper still and already outperforms the AR variant.

Table 9: Decoding cost per image (16×16 tokens) for 256-step autoregressive decoding, as in DINOSAUR and SPOT, and for the SlotMask masked decoder. Latency is measured on a single NVIDIA RTX PRO 6000 Blackwell GPU. Mask quality is on MS COCO; the autoregressive row is the AR variant of our decoder from Table 3.

Decoding	**Passes**	**GFLOPs**	**Latency (ms)**	$\text{mBO}_i \uparrow$	$\text{mBO}_c \uparrow$
Autoregressive	256	$\approx$5,000	$\approx$90	35.0	46.4
Masked, 1 step	1	$\approx$20	$\approx$0.3	35.2	47.7
Masked, 64 steps (default)	64	$\approx$500	$\approx$7	**36.1**	**48.8**

A.5 Sensitivity to the number of slots

The main experiments use the slot counts of prior work to keep comparisons matched. Table 10 sweeps K on MS COCO for SlotMask and SPOT_{v3}. SlotMask improves monotonically in both metrics as K grows, whereas for SPOT_{v3} only mBO_i improves and mBO_c peaks at $K = 7$ and degrades for larger slot budgets. SlotMask is thus more robust to the choice of K.

Table 10: Slot-count sweep on MS COCO.

	SlotMask		$\textbf{SPOT}_{\text{v3}}$	
K	$\text{mBO}_i \uparrow$	$\text{mBO}_c \uparrow$	$\text{mBO}_i \uparrow$	$\text{mBO}_c \uparrow$
5	34.2	46.2	34.3	43.4
6	35.3	47.8	35.5	44.5
7	36.1	48.8	35.3	45.3
9	37.1	49.6	36.4	44.2
11	37.7	50.3	37.1	44.5

A.6 Slot utilization

Slot-based methods often suffer from redundant slot activation. On the MS COCO validation set, we count a slot as *active* if it is assigned more than 1% of the image patches, and compute the assignment entropy $H = -\sum_k p_k \log p_k$, where p_k is the fraction of patches assigned to slot k. SlotMask uses 4.98 active slots on average with an entropy of 1.124, whereas SPOT_{v3} uses 6.31 active slots with an entropy of 1.533. Together with its stronger mask metrics, this indicates that SlotMask uses its slots more compactly and is less prone to activating redundant slots.

A.7 SLOT-AWARE INITIALIZATION AND REPRESENTATION QUALITY

To test whether slot-aware initialization improves the slot representations themselves rather than only the decoder attention maps, we repeat the property-prediction experiment of Table 2 on MS COCO for a SlotMask model trained without it (Table 11). Removing slot-aware initialization reduces label accuracy from 80.8% to 76.6% and degrades coordinate MSE, placing the model between SPOT_{v3} and the full SlotMask. Slot-aware initialization therefore improves the learned representations, not merely the masks.

Table 11: Property prediction on MS COCO with and without slot-aware initialization (MSE $\times 10^{-2}$).

Model	Acc (%) ↑	MSE ↓
SPOT_{v3}	74.3	**2.5**
SlotMask w/o slot-aware init	76.6	2.7
SlotMask	**80.8**	**2.5**