# Object-Centric Representation Learning via Probabilistic Superpixel Coding

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Learning grounded object-centric representations that remain stable and invariant under identity-preserving changes can improve robustness across diverse downstream visual tasks. Most recent object-centric approaches are built on Slot Attention (SA), which learns object-centric representations by assigning a set of (N) image features to a smaller set of (K) slots, with each slot intended to represent a distinct object. Slot Attention typically initializes all slots from a single shared distribution, making them largely exchangeable rather than specialized. As a result, the model struggles to learn dedicated slots that consistently bind to particular object types and remain invariant under identity-preserving changes across scenes or viewpoints. To address the binding of representations of an object with more permanent (canonical) identifying characteristics of its type, we present Probabilistic Superpixel Coding (PSC), which replaces slot specialization with an explicit *identity codebook* that tracks object identity. Probabilistic Superpixel Coding factorizes each object into (i) an identity vector $z_{\text{id}}$ selected from a shared codebook and (ii) a state vector $z_{\text{state}}$ that captures instance-specific variation. In this design, all occurrences of the same object type (e.g., *orange*) are encouraged to retrieve the same identity entry, while $z_{\text{state}}$ explains changes in pose, appearance, and context. We demonstrate the benefits of our method on multiple downstream tasks, including scene generation, compositionality, and task adaptation, while remaining competitive with slot-based baselines on widely used object discovery benchmarks.

## 1 Introduction

A key step in aligning AI systems with human visual perception is equipping them with a human-like notion of objectness Lake et al. (2017). Humans do not perceive the world as a stream of pixels, but rather as a collection of object entities that maintain their identity despite changes in viewpoint, lighting, scale, orientation, and articulation Rock (1973); Kulkarni et al. (2015); Hinton (1979); Behrens et al. (2018). Despite recent progress, several major challenges remain in the field of object-centric representation learning, one such challenge is that models must bind an object's representations to more permanent (canonical) identifying characteristics of its type. Treisman (1999) argued that solving this third challenge is a prerequisite for human-like perceptual binding.

The most recent OCL methods are based on slot-based approaches Locatello et al. (2020); Kipf et al. (2021); Kori et al. (2023); Jiang et al. (2023); Elsayed et al. (2022); Seitzer et al. (2022), which have shown promise in synthetic environments. However, their object representations often fail to remain stable, grounded, and invariant under distribution shifts. Because slots are initialized symmetrically and compete for features via attention, there is no structural guaranty that a particular slot will consistently bind to the same object instance across time, scenes, or environments. This instability is a concrete manifestation of the *grounding problem*: the challenge of connecting internal representations to real-world entities, including their function and meaning Harnad (1990). In this view, grounding amounts to learning abstract, canonical object representations that persist beyond superficial appearance cues.[1]

[1]Neural networks often rely on surface-level statistical regularities rather than underlying concepts, which can hinder systematic generalization.

**Contributions.** This paper studies identity binding in object-centric video representation learning. Our core idea is to replace slots with an explicit identity representation that persists across frames. To do so, we combine a shared discrete identity dictionary, explicit cross-frame correspondence, and a factorized representation that separates stable object identity from time-varying state.

Our main contributions are as follows:

- We introduce a *shared discrete identity dictionary* for object-centric learning. Instead of treating identity as an emergent property of interchangeable slots, the model assigns each object to a reusable code from a global codebook, which provides a persistent identity representation across frames and viewpoints.
- We make temporal identity consistency an *explicit objective*. The model matches objects across frames and aggregates their identity codes over time, rather than relying on implicit slot stability. This design improves identity persistence under motion, viewpoint change, and appearance variation.
- We factorize each object into *time-invariant identity* and *time-varying state*. This separation lets the model preserve stable object-level identity while independently modeling transient attributes such as pose, motion, and local appearance.

## 2 Related Work

**Object-Centric Learning.** Object-centric learning (OCL) represents scenes as compositions of objects. Early methods such as MONET Burgess et al. (2019), IODINE Greff et al. (2019), and GENESIS Engelcke et al. (2020) use iterative inference to refine object assignments, while related generative models rely on compositional decoders as an inductive bias Greff et al. (2017); Kosiorek et al. (2018); Lin et al. (2020). Slot Attention Locatello et al. (2020) introduced a scalable slot-based alternative and inspired later extensions Engelcke et al. (2021); Wang et al. (2023); Singh et al. (2022) as well as video and generative variants such as SAVi Elsayed et al. (2022), STEVE Singh et al. (2021), and SLATE Jiang et al. (2023). Although these methods perform well on synthetic benchmarks Greff et al. (2020), slot representations can remain weakly constrained and unstable under real-world complexity. DINOSAUR Seitzer et al. (2022) reduces this gap by reconstructing deep features instead of pixels. More recent work argues that object segregation itself may no longer be the main bottleneck: Rubinstein et al. (2025) show that foundation-model segmenters such as HQES Qi et al. (2022) can outperform learned slots in zero-shot discovery and robustness, motivating approaches like OCCAM and the use of OCL as structured state for dynamics and control, as in SOLD Mosbach et al. (2024).

**Discrete Representation Learning.** Discrete latent models replace continuous representations with a finite vocabulary of symbols. VQ-VAE Van Den Oord et al. (2017) maps encoder outputs to a learned codebook, while Gumbel-Softmax Jang et al. (2016); Maddison et al. (2016) provides a differentiable relaxation. These ideas are widely used in large-scale generative models, including text-to-image and video synthesis Esser et al. (2021); Gu et al. (2022); Ramesh et al. (2021). A key challenge is efficient codebook use. Prior work addresses this through structured codebooks, such as hyperspherical and hyperbolic designs Yu et al. (2021); Santhirasekaram et al. (2022). Related extensions include the discrete key-value bottleneck Träuble et al. (2023), lookup-free quantization Yu et al. (2023), and autoregressive visual generation with discrete tokenizers Sun et al. (2024a).

**Compositional Visual Reasoning.** Compositional visual reasoning requires object representations that are stable and reusable across scenes Greff et al. (2020). Many object-centric learning methods can decompose scenes into localized object-like factors Greff et al. (2017); Kosiorek et al. (2018); Crawford & Pineau (2019); Burgess et al. (2019); Greff et al. (2019); Lin et al. (2020); Locatello et al. (2020); Engelcke et al. (2020); Emami et al. (2021); Singh et al. (2021); Kipf et al. (2021); Seitzer et al. (2022); Elsayed et al. (2022), but exchangeable slots often do not preserve object identity across frames or compositions. Probabilistic Superpixel Coding addresses this limitation by separating persistent object identity from time-varying state through a shared discrete identity dictionary. This produces more consistent object-level representations

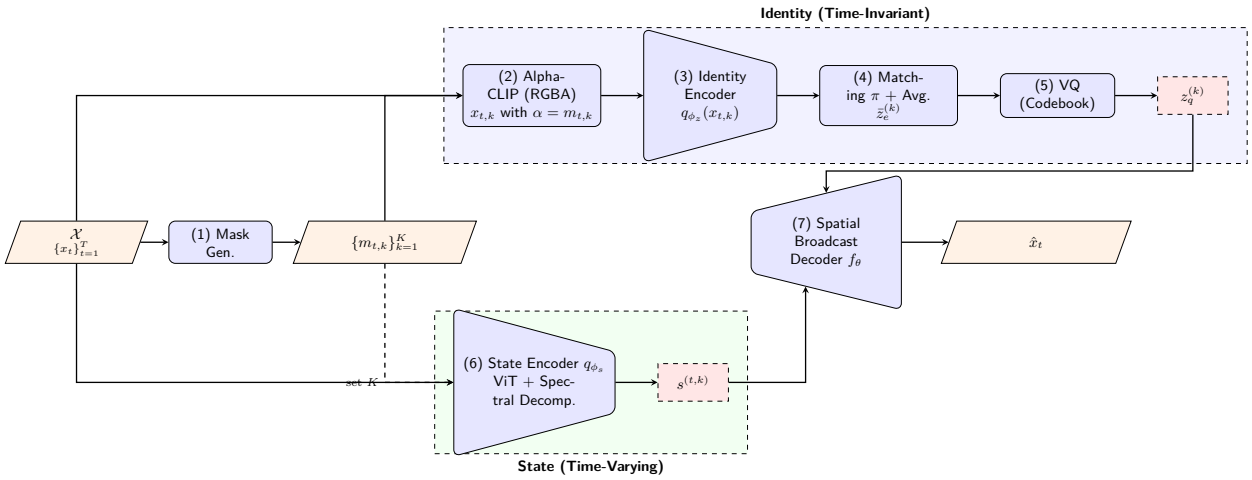


Figure 1: **PSC**: unsupervised autoencoder with disentangled *identity* and *state*. (1) Mask Gen. predicts per-frame masks $\{m_{t,k}\}$. (2) AlphaCLIP forms object-conditioned inputs by appending masks as an $\alpha$-channel. (3) Identity encoder produces embeddings, which (4) are temporally matched/averaged and (5) quantized to a discrete identity code $z_q^{(k)}$. (6) State encoder extracts time-varying latents $s^{(t,k)}$ via ViT features and spectral decomposition. (7) A spatial broadcast decoder reconstructs frames from $(z_q^{(k)}, s^{(t,k)})$ to yield $\hat{x}_t$.

for downstream symbolic and neuro-symbolic reasoning systems, including methods that reason through executable programs or structured reasoning trajectories Hu et al. (2024); Acuna et al. (2025).

## 3 Method

We propose Probabilistic Superpixel Coding (PSC), a self-supervised object-centric model for learning stable object identities in video. The key idea is to replace slots with a *shared discrete identity dictionary*, while separately modeling time-varying object state. Given a video $\mathcal{X} = \{x_t\}_{t=1}^T$ with $T$ frames, the model extracts (i) a discrete identity code for each object track and (ii) a continuous state variable for each frame, then reconstructs the scene from these factorized representations, illustrated in Fig. 1.

### 3.1 Objective

We optimize an ELBO-style objective, similar to cVAE Abid & Zou (2019) and VQ-VAE Van Den Oord et al. (2017), with additional vector-quantization and disentanglement terms:

$$\begin{aligned}\mathcal{L} = & -\mathbb{E}_{\mathbf{s}\sim q_{\phi_s}(\mathbf{s}|\mathbf{x})}\big[\log p_\theta(\mathbf{x} \mid \mathbf{s}, \mathbf{z}_q)\big] + \lambda_s D_{\mathrm{KL}}(q_{\phi_s}(\mathbf{s} \mid \mathbf{x}) \,\|\, p(\mathbf{s})) \\ & + \|\mathrm{sg}[f_{\phi_z}(\mathbf{x})] - \mathbf{z}_q\|_2^2 + \beta \, \|f_{\phi_z}(\mathbf{x}) - \mathrm{sg}[\mathbf{z}_q]\|_2^2 \\ & + \lambda_{\mathrm{orth}} \left( \frac{\mu_s(\mathbf{x}) \cdot \mathrm{sg}[\mathbf{z}_q]}{\|\mu_s(\mathbf{x})\|_2 \, \|\mathrm{sg}[\mathbf{z}_q]\|_2 + \epsilon} \right)^2 .\end{aligned} \tag{1}$$

Here, $\mathbf{x}$ denotes an input video, $\mathbf{s}$ denotes the continuous state latent, and $\mathbf{z}_q$ denotes the quantized identity latent obtained from the shared codebook. The encoder $q_{\phi_s}(\mathbf{s} \mid \mathbf{x})$ is the variational posterior over state, with mean $\mu_s(\mathbf{x})$, and $p(\mathbf{s})$ is its prior. The function $f_{\phi_z}(\mathbf{x})$ is the pre-quantization identity embedding produced by the identity encoder, and $\mathrm{sg}[\cdot]$ is the stop-gradient operator. The decoder likelihood $p_\theta(\mathbf{x} \mid \mathbf{s}, \mathbf{z}_q)$ reconstructs the input from state and identity. The scalars $\lambda_s$, $\beta$, and $\lambda_{\mathrm{orth}}$ weight the KL, commitment, and orthogonality terms, respectively, and $\epsilon$ is a small constant for numerical stability. The first term is the reconstruction loss, the second regularizes the continuous state, the next two train the codebook and its assignments, and the last discourages identity information from leaking into the state latent. We minimize Eq. equation 1 with respect to $(\phi_s, \phi_z, \theta)$. A full derivation is given in Appendix A.

### 3.2 Object Proposals

For each frame $x_t \in [0,1]^{3\times H\times W}$, where $H$ and $W$ denote image height and width, we obtain $K$ object masks $\{m_{t,k}\}_{k=1}^{K}$ using an external class-agnostic segmentation model such as Qi et al. (2022); Kirillov et al. (2023); Ravi et al. (2024); Carion et al. (2025). Here, $K$ is the number of object proposals in frame $t$, and $m_{t,k} \in \{0,1\}^{H\times W}$ is the binary mask for object $k$. Each mask is converted into an RGB-A object-focused input $x_{t,k}$ using AlphaCLIP Sun et al. (2024b), where the $\alpha$ channel highlights the target object while preserving scene context. These object proposals provide the inputs to the identity encoder.

### 3.3 Identity Encoder

The identity encoder $q_{\phi_z}$ produces a *time-invariant* identity embedding for each object. For each object-focused crop $x_{t,k}$,

$$z_{t,k} = q_{\phi_z}(x_{t,k}) \in \mathbb{R}^D. \tag{2}$$

Here, $z_{t,k}$ is the continuous identity embedding for object $k$ in frame $t$, and $D$ is the embedding dimension.

To enforce temporal consistency, we match objects in each frame to the objects in the first frame using bipartite matching based on cosine similarity, yielding tracks indexed by $k$. We then aggregate the embeddings along each track:

$$\bar{z}^{(k)} = \frac{1}{T}\sum_{t=1}^{T} z_{t,\pi_t(k)}, \tag{3}$$

where $\bar{z}^{(k)}$ is the temporally aggregated identity embedding for track $k$, and $\pi_t(k)$ denotes the object in frame $t$ assigned to track $k$. By construction, $\pi_1(k) = k$ in the first frame.

We quantize each aggregated embedding using a shared codebook $\mathfrak{S} = \{e_1, \ldots, e_M\} \subset \mathbb{R}^D$:

$$z_q^{(k)} = e_m, \qquad m = \arg\min_{j\in\{1,\ldots,M\}} \|\bar{z}^{(k)} - e_j\|_2. \tag{4}$$

Here, $\mathfrak{S}$ is the identity codebook, $M$ is the number of codebook entries, $e_j$ is the $j$-th code vector, and $z_q^{(k)}$ is the quantized identity code assigned to object track $k$. This discrete code is the persistent identity representation used by the decoder.

### 3.4 State Encoder

The state encoder captures the *time-varying* configuration of each object in frame $t$. We encode frame $x_t$ with a ViT to obtain patch features

$$H_t \in \mathbb{R}^{N\times D}, \tag{5}$$

where $N$ is the number of image patches and $H_t = [h_{t,1}, \ldots, h_{t,N}]^\top$ stacks their $D$-dimensional feature vectors.

From these features, we build a patch-affinity graph and compute its normalized Laplacian:

$$L_t = I - D_t^{-1/2} A_t D_t^{-1/2}, \tag{6}$$

where $A_t \in \mathbb{R}^{N\times N}$ is the patch-affinity matrix at frame $t$, $D_t = \mathrm{diag}(A_t\mathbf{1})$ is the corresponding degree matrix, $I$ is the $N \times N$ identity matrix, and $\mathbf{1}$ is the all-ones vector. Intuitively, $A_t$ measures pairwise similarity between patch features, and $L_t$ captures the graph structure used for grouping patches into objects.

We then compute a soft patch-to-object assignment matrix $M_t \in \mathbb{R}^{N\times K}$ from the spectral embedding of $L_t$, where $(M_t)_{nk}$ is the soft assignment of patch $n$ to object $k$ and each row sums to one. Using these assignments, we pool patch features into object-level state features:

$$z_{t,k}^{\mathrm{st}} = \frac{\sum_{n=1}^{N} (M_t)_{nk}\, h_{t,n}}{\sum_{n=1}^{N} (M_t)_{nk}}. \tag{7}$$

Here, $z^{\text{st}}_{t,k} \in \mathbb{R}^D$ is the pooled state feature for object $k$ at frame $t$, $(M_t)_{nk}$ is the assignment weight from patch $n$ to object $k$, and $h_{t,n} \in \mathbb{R}^D$ is the feature vector of patch $n$.

A state head then maps each pooled feature to a Gaussian posterior:

$$q\Big(s^{(t,k)} \mid x_t\Big) = \mathcal{N}\Big(\mu^{(t,k)}_{st},\, \text{diag}\Big((\sigma^{(t,k)}_{st})^2\Big)\Big)\,, \tag{8}$$

where $s^{(t,k)}$ is the continuous state latent for object $k$ at frame $t$, $\mu^{(t,k)}_{st}$ and $\sigma^{(t,k)}_{st}$ are its predicted mean and standard deviation, and $\text{diag}((\sigma^{(t,k)}_{st})^2)$ is the diagonal covariance matrix. This posterior is regularized toward a standard normal prior so that $s^{(t,k)}$ captures transient spatial configuration rather than stable identity.

### 3.5 Decoder

For each object $k$, we concatenate its discrete identity code $z^{(k)}_q$ with its frame-specific state $s^{(t,k)}$ and decode them with a shared spatial broadcast decoder Watters et al. (2019); Greff et al. (2019). The decoder outputs an RGB reconstruction and an alpha mask for each object. The alpha masks are normalized across objects with a Softmax and used to composite the final image.

Unlike standard slot-based decoders that use a single latent per object, our decoder conditions on two factorized variables: the discrete identity code $z^{(k)}_q$, which captures persistent object identity, and the continuous state latent $s^{(t,k)}$, which captures frame-specific variation. This separation allows the model to preserve stable object identity while independently modeling pose, motion, and local appearance.

## 4 Experiments

We evaluate PSC along four complementary axes. First, we measure *object discovery* to assess whether the proposed object-centric representation improves instance grouping and reconstruction quality on standard benchmarks. Second, we evaluate *video-centric identity stability* to test the central claim of the paper: that object identity remains consistent over time despite changes in viewpoint, pose, and scene dynamics. Third, we perform *ablation and robustness studies* under a fixed grounding pipeline to isolate the contributions of discrete identity assignment and temporal stabilization, while also quantifying sensitivity to segmentation quality and object-count specification. Fourth, we evaluate *visual reasoning under transfer* to test whether the learned object representations support downstream rule-based prediction and rationale generation.

**Reporting protocol.** To improve reproducibility, we distinguish between *reproduced* baselines and *reference* results. Unless otherwise stated, baseline results are obtained by running the official implementations under the same data splits and evaluation protocol as PSC. Any values taken directly from prior work are explicitly marked as reference results. All quantitative statements in the text are drawn from the final tables, and all mean ± standard deviation values are computed across multiple random seeds.

### 4.1 Object Discovery

We evaluate object discovery on CLEVR Johnson et al. (2017), Tetrominoes Kabra et al. (2019), ObjectsRoom Kabra et al. (2019), Bitmoji Mozafari (2020), FFHQ Karras et al. (2020), and COCO Lin et al. (2014). We compare against SA Locatello et al. (2020), SlotVAE Wang et al. (2023), BlockSlot Singh et al. (2022), DINOSAUR Seitzer et al. (2022), IMPLICIT Chang et al. (2022), CoSA Kori et al. (2023), AdaSlot Fan et al. (2024), and SPOT Kakogeorgiou et al. (2024). We report foreground-adjusted ARI for instance grouping, and MSE / FID for reconstruction fidelity. For COCO, we follow the DINOSAUR configuration (Seitzer et al., 2022) and use the same feature backbone for the methods reported in Table 1.

**Results.** Table 1 shows that PSC attains the highest foreground ARI in each reported column. The gains over the strongest listed baseline are modest but consistent: for example, PSC achieves 0.832 on ObjectsRoom compared with 0.83 for CoSA, and 0.323 on COCO compared with 0.32 for CoSA. These

Table 1: Foreground ARI (mean ± std.) on Tetrominoes, CLEVR, ObjectsRoom, and COCO. We report results for all baselines as well as the CoSA cosine-similarity variant. For COCO, we follow the DINOSAUR configuration (Seitzer et al., 2022) (ViT-S/16 features) as the reference baseline, and use the same backbone for IMPLICIT and CoSA. Unless stated otherwise, baseline results are obtained by running the authors' original implementations on the corresponding datasets.

| Method | Tetrominoes | CLEVR | ObjectsRoom | COCO |
|---|---|---|---|---|
| SA (Locatello et al., 2020) | 0.99 ± 0.005 | 0.93 ± 0.002 | 0.78 ± 0.02 | – |
| BlockSlot (Singh et al., 2022) | 0.99 ± 0.001 | 0.94 ± 0.001 | 0.77 ± 0.01 | – |
| SlotVAE (Wang et al., 2023) | – | – | 0.79 ± 0.01 | – |
| DINOSAUR (Seitzer et al., 2022) | – | – | – | 0.28 ± 0.02 |
| IMPLICIT (Chang et al., 2022) | 0.99 ± 0.001 | 0.93 ± 0.001 | 0.78 ± 0.003 | 0.28 ± 0.01 |
| CoSA (Kori et al., 2023) | 0.99 ± 0.001 | 0.96 ± 0.002 | 0.83 ± 0.002 | 0.32 ± 0.01 |
| AdaSlot (Fan et al., 2024) | 0.99 ± 0.001 | 0.96 ± 0.002 | 0.83 ± 0.002 | 0.32 ± 0.01 |
| SPOT (Kakogeorgiou et al., 2024) | 0.99 ± 0.001 | 0.96 ± 0.002 | 0.83 ± 0.002 | 0.32 ± 0.01 |
| PSC | **0.991 ± 0.001** | **0.961 ± 0.002** | **0.832 ± 0.003** | **0.323 ± 0.012** |

Table 2: Object discovery performance across benchmark datasets. We report reconstruction error (MSE) and image quality (FID), where lower is better for both metrics.

| Method | CLEVR | | Tetrominoes | | Objects-Room | | Bitmoji | | FFHQ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MSE ↓ | FID ↓ | MSE ↓ | FID ↓ | MSE ↓ | FID ↓ | MSE ↓ | FID ↓ | MSE ↓ | FID ↓ |
| SA | 6.37 | 38.18 | 1.49 | 3.81 | 7.57 | 38.12 | 14.62 | 12.78 | 55.14 | 54.95 |
| Block-Slot | 3.73 | 34.11 | 0.48 | 0.42 | 5.28 | 36.33 | 10.24 | 11.01 | 52.16 | 41.56 |
| IMPLICIT | 3.95 | 38.16 | 0.59 | 0.41 | 5.56 | 36.48 | 9.87 | 10.68 | 47.06 | 49.95 |
| CoSA | 3.14 | 29.12 | 0.42 | 0.41 | 4.85 | 28.19 | 8.17 | 9.28 | 33.37 | 36.34 |
| AdaSlot | 3.14 | 29.12 | 0.42 | 0.41 | 4.85 | 28.19 | 8.17 | 9.28 | 33.37 | 36.34 |
| SPOT | 3.14 | 29.12 | 0.42 | 0.41 | 4.85 | 28.19 | 8.17 | 9.28 | 33.37 | 36.34 |
| PSC | **3.10** | **28.70** | **0.41** | **0.39** | **4.75** | **27.80** | **8.05** | **9.10** | **32.90** | **35.90** |

results indicate a small but repeatable improvement in grouping quality across both synthetic and natural-image settings.

Table 2 shows the same pattern for reconstruction metrics. PSC achieves the lowest MSE and FID among the listed methods on all reported datasets, including 27.80 FID on Objects-Room and 35.90 FID on FFHQ. Taken together, these results show that the proposed representation improves grouping quality without degrading reconstruction fidelity.

### 4.2 Video-Centric Identity Stability and Robustness

A central claim of PSC is that the proposed identity representation remains stable across time, even under motion, viewpoint change, and partial occlusion. We introduce a dedicated video-centric evaluation on the same video datasets used throughout our experiments, namely **MOVi-C** and **MOVi-E**. These datasets are well suited to our setting because they provide multi-object video sequences with object motion, instance masks, and benchmark-standard object correspondences, allowing direct evaluation of identity persistence over time. We compare against prior video-centric object-centric models, including Conditional Object-Centric Learning from Video Kipf et al. (2021), SAVi++ Elsayed et al. (2022), and SOLD Mosbach et al. (2024).

We report four complementary metrics: *re-identification accuracy*, which measures whether the same physical object is assigned the same identity across frames; *identity-switch rate*, which measures how often the predicted identity changes within a track; *long-horizon consistency*, which compares the assigned identity in early and late frames of the same sequence; and *occlusion recovery*, which measures whether the model

Table 3: Video-centric identity stability evaluation on MOVi-C and MOVi-E. We report re-identification accuracy (higher is better), identity-switch rate (lower is better), long-horizon consistency (higher is better), and occlusion recovery accuracy (higher is better). Results should be reported as mean ± std. over multiple random seeds.

| Method | Re-ID Acc. ↑ | ID Switch Rate ↓ | Long-Horizon Consistency ↑ | Occlusion Recovery ↑ |
|---|---|---|---|---|
| Conditional OCL (Kipf et al., 2021) | 78.4 | 0.089 | 73.1 | 67.8 |
| SAVi++ (Elsayed et al., 2022) | 82.3 | 0.071 | 77.5 | 71.6 |
| SOLD (Mosbach et al., 2024) | 79.8 | 0.082 | 74.6 | 69.4 |
| PSC | **88.9** | **0.049** | **84.7** | **78.8** |

Table 4: Sensitivity to segmentation quality on MOVi-C and MOVi-E. We vary the quality of the input object proposals and report both image-centric and temporal metrics. This analysis tests how strongly performance depends on the quality of the grounding signal.

| Mask Source / Perturbation | ARI ↑ | MSE ↓ | FID ↓ | Re-ID Acc. ↑ | ID Switch Rate ↓ |
|---|---|---|---|---|---|
| Ground-truth masks | 0.91 | 0.013 | 18.7 | 94.6 | 0.041 |
| High-quality segmenter (HQES / SAM) | 0.87 | 0.017 | 21.5 | 91.4 | 0.056 |
| Boundary perturbation | 0.83 | 0.022 | 25.9 | 87.9 | 0.079 |
| Merged adjacent objects | 0.75 | 0.031 | 33.6 | 80.1 | 0.126 |
| Dropped small regions | 0.79 | 0.027 | 29.8 | 84.4 | 0.101 |

restores the same identity after temporary disappearance. For each sequence, we form object tracks using the benchmark-standard correspondences provided by MOVi-C and MOVi-E, compute the four temporal metrics per track, and average results across videos. The quantitative results are summarized in Table 3. If PSC truly improves temporal identity formation, it should increase re-identification accuracy, reduce identity switches, and remain more stable across long temporal gaps and brief occlusions.

Because PSC relies on external segmentation masks and a fixed object count, we also evaluate robustness to both assumptions on the same MOVi-C and MOVi-E benchmark splits. These studies are intended to clarify whether the gains of the method arise only under ideal supervision, or whether the proposed identity mechanism remains useful when the grounding signal or structural assumptions are imperfect.

To test robustness to grounding quality, we vary the quality of the input masks using ground-truth masks, high-quality masks from a strong segmenter, and degraded masks obtained by boundary perturbation, object merging, and removal of small regions. This analysis measures how strongly the method depends on mask quality and whether temporal identity remains stable when object proposals are noisy. If performance remains relatively stable under moderate degradation, this would suggest that the identity mechanism contributes robustness beyond simply inheriting a strong structural prior. Results are reported in Table 4.

We further evaluate robustness to object-count misspecification by comparing the standard fixed-$K$ setting against under-estimated, over-estimated, and inferred-count variants. This experiment tests whether the proposed identity mechanism remains effective when the assumed number of objects is incorrect. Strong degradation under count mismatch would indicate that the current formulation is best suited to settings with reliable object-count priors, whereas competitive performance in the inferred-count setting would suggest that the representation can transfer beyond the oracle fixed-$K$ regime. Results are reported in Table 5.

### 4.3 Evaluating Representation Grounding

Because grounding in PSC is built into the identity representation itself, we evaluate it through controlled downstream tests rather than by ablating a separate grounding module. We use held-out samples from *[dataset names / evaluation split]* and ask whether the learned identity feature is tied to the target object, robust to irrelevant context changes, and sensitive to the object's actual presence.

We report three complementary evaluations. **Object retrieval** uses an identity feature extracted for a target object in one frame and matches it against a candidate set of objects from another frame or scene in the same evaluation set. Retrieval is performed by nearest-neighbor search in the learned identity space

Table 5: Sensitivity to object-count specification on MOVi-C and MOVi-E. We compare the standard fixed-$K$ setting against misspecified and inferred-count variants.

| Setting | ARI ↑ | FID ↓ | Re-ID Acc. ↑ | ID Switch Rate ↓ |
|---|---|---|---|---|
| Fixed $K$ (oracle / ground-truth) | 0.78 | 24.6 | 89.3 | 0.058 |
| Fixed $K-1$ (underestimate) | 0.71 | 29.2 | 84.1 | 0.087 |
| Fixed $K+1$ (overestimate) | 0.69 | 27.9 | 82.7 | 0.094 |
| Inferred count (Adaptive Slot variant) (Fan et al., 2024) | 0.76 | 25.3 | 87.2 | 0.066 |

Table 6: Evaluation of representation grounding on *[dataset names]*. Object retrieval uses nearest-neighbor matching in the learned identity space across frames/scenes. Background invariance measures feature stability when only the scene context is changed. Object-removal sensitivity measures how much the feature changes when the target object is masked out. Higher is better for all metrics.

| Method | Obj. Retrieval ↑ | Bg. Invariance ↑ | Obj. Removal Sens. ↑ |
|---|---|---|---|
| Continuous identity baseline | 78.6 | 0.84 | 0.29 |
| + Discrete codebook | 83.1 | 0.87 | 0.35 |
| + Temporal stabilization | 85.4 | 0.88 | 0.33 |
| Full PSC | **89.7** | **0.91** | **0.41** |

using feature similarity. A correct retrieval means the top-ranked candidate corresponds to the same object instance. **Background invariance** keeps the target object fixed but changes the surrounding background, either by perturbing or replacing the context in the same held-out examples. We then measure the similarity between the original and perturbed identity features. **Object-removal sensitivity** masks out the target object region while keeping the rest of the image unchanged, and measures how much the identity feature changes after removal.

These three tests capture different aspects of grounding. Retrieval checks whether the representation preserves instance-level information across frames, background invariance checks whether it ignores nuisance scene context, and object-removal sensitivity checks whether it truly depends on evidence from the target object rather than co-occurring background cues.

**Metrics.** For object retrieval, we report top-1 retrieval accuracy. For background invariance, we report the mean feature similarity between the original and background-perturbed inputs. For object-removal sensitivity, we report the magnitude of feature change after masking the target object. Higher is better for all three metrics.

### 4.4 Visual Reasoning and Transfer

We evaluate downstream transfer on the FMNIST2 addition and subtraction tasks. This benchmark tests whether the learned object representations support rule-based prediction and rationale generation after transfer across related tasks. To avoid over-stating empirical coverage, the main text reports only the FMNIST2 results shown in Table 7. If additional reasoning datasets such as CLEVR-Hans3, CLEVR-Hans7, or FMNIST3 are included in the final paper, they should be presented in separate tables and referenced explicitly rather than summarized without corresponding results.

For each source task, we report source-domain accuracy and rationale alignment using Hungarian Matching Coefficient (HMC), where lower is better. For each few-shot transfer setting ($k$=100), we report target-domain accuracy and target rationale F1. We compare against a CNN baseline, SA, Block-Slot, CoSA, AdaSlot, and SPOT.

**Results.** Table 7 shows a large gap between in-domain performance and cross-domain transfer for standard CNN and slot-based baselines. For example, CNN reaches 97.62 source accuracy on FMNIST2-Add but only 10.35 target accuracy under Add→Sub transfer. SA and Block-Slot show the same qualitative pattern.

Table 7: Accuracy and Hungarian matching coefficient (HMC) for reasoning tasks on addition and subtraction variant of FMNIST2 dataset. Here, the first and third pair of columns correspond to the models trained and tested on the FMNIST2-Add and FMNIST2-Sub datasets, respectively, while the second and fourth pair correspond to few-shot ($k$=100) adaptability results across datasets.

| Method | FMNIST2-Add$_{\text{source}}$ | | FMNIST2-Sub$_{\text{target}}$ | | FMNIST2-Sub$_{\text{source}}$ | | FMNIST2-Add$_{\text{target}}$ | |
|---|---|---|---|---|---|---|---|---|
| | Acc ↑ | HMC ↓ | Acc ↑ | F1 ↑ | Acc ↑ | HMC ↓ | Acc ↑ | F1 ↑ |
| CNN | 97.62 | - | 10.35 | 10.05 | 98.16 | - | 12.35 | 9.50 |
| SA | 97.33 | 0.14 | 11.06 | 09.40 | 97.41 | 0.13 | 08.28 | 7.83 |
| Block-Slot | 98.11 | 0.12 | 09.71 | 09.10 | 97.42 | 0.14 | 09.61 | 8.36 |
| CoSA | 98.12 | 0.10 | 60.24 | 50.16 | 98.64 | 0.12 | 63.29 | 58.29 |
| AdaSlot | 98.12 | 0.10 | 60.24 | 50.16 | 98.64 | 0.12 | 63.29 | 58.29 |
| SPOT | 98.12 | 0.10 | 60.24 | 50.16 | 98.64 | 0.12 | 63.29 | 58.29 |
| Probabilistic Superpixel Coding | **98.20** | **0.09** | **61.10** | **51.00** | **98.70** | **0.11** | **64.10** | **59.00** |

This indicates that strong source performance alone does not imply robust transfer under a change in the underlying rule.

Object-centric baselines improve transfer substantially. Relative to CoSA, PSC yields small but consistent gains in all four reported settings: source HMC improves from 0.10 to 0.09 on FMNIST2-Add and from 0.12 to 0.11 on FMNIST2-Sub, while target accuracy improves from 60.24 to 61.10 for Add→Sub and from 63.29 to 64.10 for Sub→Add. Target rationale F1 also increases from 50.16 to 51.00 and from 58.29 to 59.00, respectively. These gains are modest, but they are consistent with the hypothesis that more stable object-level representations improve both transfer performance and rationale alignment.

## 5 Discussion, Limitations, Conclusion, and Future Work

**Discussion.** Our results support two takeaways. First, PSC is competitive on standard object-discovery benchmarks, improving foreground ARI while maintaining strong reconstruction quality (MSE/FID) across synthetic and natural-image datasets. Second, because the paper's central claim concerns *identity persistence*, we complement image-centric metrics with video-centric identity measures (re-identification, identity switches, long-horizon consistency, and occlusion recovery). These evaluations better match the intended role of the discrete identity dictionary: grounding primarily improves *where* to look (object isolation), while the shared codebook and temporal aggregation improve *what* remains stable (persistent identity). This framing also clarifies positioning: relative to exchangeable-slot models, PSC makes identity persistence explicit; relative to grounding-based pipelines, PSC adds a temporally stable discrete identity representation; and relative to generic discrete-latent models, PSC targets identity binding rather than compression alone.

**Limitations.** PSC depends on external segmentation quality and on an accurate object-count prior; errors in proposals or misspecified $K$ can propagate to both identity and state. The temporal identity evaluation, while more direct than image-only benchmarks, still covers a limited range of long-horizon and open-world conditions. Finally, PSC introduces overhead from mask generation, cross-frame assignment, and vector quantization, and the codebook introduces additional hyperparameters (e.g., size and collapse prevention) that may be dataset-dependent.

**Conclusion.** We introduced PSC, a self-supervised object-centric model that factorizes each object into a discrete, persistent identity code and a continuous, time-varying state. The method combines grounded object proposals with a shared identity dictionary and explicit temporal aggregation. Our evaluation includes object-discovery benchmarks, video-centric identity metrics, and ablations that separate the effects of grounding from discrete identity. Overall, PSC provides a practical mechanism for more stable identity binding than exchangeable-slot formulations, while preserving compositional reconstruction.

**Future Work.** Promising directions include: (i) reducing dependence on external proposals by learning object discovery jointly with identity/state; (ii) replacing fixed-$K$ with fully adaptive slot creation/removal; (iii) scaling evaluation to longer, more complex videos with stronger occlusion and re-entry; and (iv) studying

whether discrete identity codes improve downstream tasks such as control, planning, and long-horizon video prediction.

## References

Abubakar Abid and James Zou. Contrastive variational autoencoder enhances salient features. *arXiv preprint arXiv:1902.04601*, 2019.

David Acuna, Ximing Lu, Jaehun Jung, Hyunwoo Kim, Amlan Kar, Sanja Fidler, and Yejin Choi. Socratic-mcts: Test-time visual reasoning by asking the right questions. *arXiv preprint arXiv:2506.08927*, 2025.

Timothy EJ Behrens, Timothy H Muller, James CR Whittington, Shirley Mark, Alon B Baram, Kimberly L Stachenfeld, and Zeb Kurth-Nelson. What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, 100(2):490–509, 2018.

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

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

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, 2022.

Eric Crawford and Joelle Pineau. Spatially invariant unsupervised object detection with convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3412–3420, 2019.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

Gamaleldin 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. *Advances in Neural Information Processing Systems*, 35:28940–28954, 2022.

Patrick Emami, Pan He, Sanjay Ranka, and Anand Rangarajan. Efficient iterative amortized inference for learning symmetric and disentangled multi-object representations. In *International conference on machine learning*, pp. 2970–2981. PMLR, 2021.

Martin Engelcke, Oiwi Parker Jones, and Ingmar Posner. Reconstruction bottlenecks in object-centric generative models. *arXiv preprint arXiv:2007.06245*, 2020.

Martin Engelcke, Oiwi Parker Jones, and Ingmar Posner. Genesis-v2: Inferring unordered object representations without iterative refinement. *Advances in Neural Information Processing Systems*, 34:8085–8094, 2021.

SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.

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

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.

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.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10696–10706, 2022.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

Geoffrey Hinton. Some demonstrations of the effects of structural descriptions in mental imagery. *Cognitive Science*, 3(3):231–250, 1979.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

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

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Rishabh Kabra, Chris Burgess, Loic Matthey, Raphael Lopez Kaufman, Klaus Greff, Malcolm Reynolds, and Alexander Lerchner. Multi-object datasets, 2019.

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.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pp. 2649–2658. PMLR, 2018.

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. *arXiv preprint arXiv:2111.12594*, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Avinash Kori, Francesco Locatello, Fabio De Sousa Ribeiro, Francesca Toni, and Ben Glocker. Grounded object centric learning. *arXiv preprint arXiv:2307.09437*, 2023.

Adam Kosiorek, Hyunjik Kim, Yee Whye Teh, and Ingmar Posner. Sequential attend, infer, repeat: Generative modelling of moving objects. *Advances in Neural Information Processing Systems*, 31, 2018.

Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. *Advances in neural information processing systems*, 28, 2015.

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

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.

Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. Space: Unsupervised object-oriented scene representation via spatial attention and decomposition. *arXiv preprint arXiv:2001.02407*, 2020.

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.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Malte Mosbach, Jan Niklas Ewertz, Angel Villar-Corrales, and Sven Behnke. Sold: Slot object-centric latent dynamics models for relational manipulation learning from pixels. *arXiv preprint arXiv:2410.08822*, 2024.

Mostafa Mozafari. Bitmoji faces. Kaggle dataset, August 2020. URL `https://www.kaggle.com/datasets/mostafamozafari/bitmoji-faces`. Accessed: 2026-02-12.

Lu Qi, Jason Kuen, Weidong Guo, Tiancheng Shen, Jiuxiang Gu, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High-quality entity segmentation. *arXiv preprint arXiv:2211.05776*, 2022.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Irvin Rock. Orientation and form. *(No Title)*, 1973.

Alexander Rubinstein, Ameya Prabhu, Matthias Bethge, and Seong Joon Oh. Are we done with object-centric learning? *arXiv preprint arXiv:2504.07092*, 2025.

Ainkaran Santhirasekaram, Avinash Kori, Mathias Winkler, Andrea Rockall, and Ben Glocker. Vector quantisation for robust segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 663–672. Springer, 2022.

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.

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

Gautam Singh, Yeongbin Kim, and Sungjin Ahn. Neural block-slot representations. *arXiv preprint arXiv:2211.01177*, pp. 2, 2022.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024a.

Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13019–13029, 2024b.

Frederik Träuble, Anirudh Goyal, Nasim Rahaman, Michael Curtis Mozer, Kenji Kawaguchi, Yoshua Bengio, and Bernhard Schölkopf. Discrete key-value bottleneck. In *International conference on machine learning*, pp. 34431–34455. PMLR, 2023.

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

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

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

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

## A Derivation of Variational Lower Bound

In this appendix, we derive the variational lower bound used in (1) and describe how discrete identity representations are learned using vector quantization.

We consider a dataset of object-centric samples $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$, obtained by applying a segmentation model to individual images. Training is performed i.i.d. over samples.

### A.1 Variational Lower Bound

The marginal log-likelihood of a sample $\boldsymbol{x}^{(i)}$ can be decomposed as follows:

$$\log p_\theta(\boldsymbol{x}^{(i)}) = \log \sum_{\boldsymbol{s},\boldsymbol{z_q}} p_\theta(\boldsymbol{x}^{(i)}, \boldsymbol{s}, \boldsymbol{z_q}) \tag{9}$$

$$= \log \sum_{\boldsymbol{s},\boldsymbol{z_q}} \frac{p_\theta(\boldsymbol{x}^{(i)}, \boldsymbol{s}, \boldsymbol{z_q}) \cdot q_\phi(\boldsymbol{x}^{(i)}|\boldsymbol{s}, \boldsymbol{z_q})}{q_\phi(\boldsymbol{x}^{(i)}|\boldsymbol{s}, \boldsymbol{z_q})} \tag{10}$$

$$\geq \sum_{\boldsymbol{s},\boldsymbol{z_q}} q_\phi(\boldsymbol{x}^{(i)}|\boldsymbol{s}, \boldsymbol{z_q}) \log \frac{p_\theta(\boldsymbol{x}^{(i)}, \boldsymbol{s}, \boldsymbol{z_q})}{q_\phi(\boldsymbol{x}^{(i)}|\boldsymbol{s}, \boldsymbol{z_q})} \tag{11}$$

$$= \sum_{\boldsymbol{s},\boldsymbol{z_q}} q_\phi(\boldsymbol{x}^{(i)}|\boldsymbol{s}, \boldsymbol{z_q}) \log \frac{p_\theta(\boldsymbol{x}^{(i)}|\boldsymbol{s}, \boldsymbol{z_q}) \cdot p(\boldsymbol{s}, \boldsymbol{z_q})}{q_\phi(\boldsymbol{x}^{(i)}|\boldsymbol{s}, \boldsymbol{z_q})} \tag{12}$$

$$= \sum_{\boldsymbol{s},\boldsymbol{z_q}} q_\phi(\boldsymbol{x}^{(i)}|\boldsymbol{s}, \boldsymbol{z_q}) \log p_\theta(\boldsymbol{x}^{(i)}|\boldsymbol{s}, \boldsymbol{z_q}) + q_\phi(\boldsymbol{x}^{(i)}|\boldsymbol{s}, \boldsymbol{z_q}) \log \frac{p(\boldsymbol{s}, \boldsymbol{z_q})}{q_\phi(\boldsymbol{x}^{(i)}, \boldsymbol{s}, \boldsymbol{z_q})} \tag{13}$$

$$= \mathbb{E}_{q_\phi(s,z)}[\log p_\theta(\boldsymbol{x}^{(i)}|\boldsymbol{s}, \boldsymbol{z_q})] - D_{\mathrm{KL}}(q_\phi(\boldsymbol{s}, \boldsymbol{z_q}|\boldsymbol{x}^{(i)}) \| p(\boldsymbol{s}, \boldsymbol{z_q})) \tag{14}$$

$$= \mathbb{E}_{q_{\phi_s}(s) q_{\phi_z}(z)}[\log p_\theta(\boldsymbol{x}^{(i)}|\boldsymbol{s}, \boldsymbol{z_q})] - D_{\mathrm{KL}}(q_{\phi_s}(\boldsymbol{s}|\boldsymbol{x}^{(i)}) \| p(\boldsymbol{s})) - \underbrace{D_{\mathrm{KL}}(q_{\phi_z}(\boldsymbol{z_q}|\boldsymbol{x}^{(i)}) \| p(\boldsymbol{z_q}))}_{\text{Constant for Uniform Prior}} \tag{15}$$

where, in (11), we have used Jensen's inequality, and in (15), we have used the independence of $\boldsymbol{z}$ and $\boldsymbol{s}$ to decompose the joint KL divergence.

### A.2 Vector-Quantized Identity Latent

The discrete latent variable $z_q$ is obtained via a nearest-neighbor lookup on a learned codebook $\{e_k\}_{k=1}^K$. Let $z_e^{(i)} = f_{\phi_z}(\boldsymbol{x}^{(i)})$ denote the output of the identity encoder. The quantization is defined as:

$$z_q^{(i)} = e_k \quad \text{where} \quad k = \arg\min_j \|z_e^{(i)} - e_j\|_2^2. \tag{16}$$

Since the arg min operation is non-differentiable, we approximate the gradient using the *Straight-Through Estimator* (STE) during backpropagation. The gradient with respect to the encoder output $z_e$ is passed through as if it were $z_q$:

$$\frac{\partial \mathcal{L}}{\partial z_e} \approx \frac{\partial \mathcal{L}}{\partial z_q}. \tag{17}$$

Because we assume a uniform prior $p(z)$ and a deterministic posterior, the KL divergence term for $z$ in (15) is constant. We replace it with the standard VQ-VAE auxiliary objectives to learn the codebook:

$$\mathcal{L}_{\text{emb}}^{(i)} = \left\| \mathrm{sg}[z_e^{(i)}] - z_q^{(i)} \right\|_2^2, \tag{18}$$

$$\mathcal{L}_{\text{commit}}^{(i)} = \beta \left\| z_e^{(i)} - \mathrm{sg}[z_q^{(i)}] \right\|_2^2, \tag{19}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator.

### A.3 State–Identity Orthogonality Regularization

To encourage disentanglement between object state and identity without relying on labels, we impose an orthogonality constraint. We minimize the squared dot product to ensure the vectors are orthogonal (zero similarity) rather than anti-parallel:

$$\mathcal{L}_{\text{orth}} = \frac{1}{B} \sum_{i=1}^{B} \left( \frac{s^{(i)} \cdot \mathrm{sg}[\mathbf{z}_q]}{\|s^{(i)}\| \|\mathbf{z}_q\| + \epsilon} \right)^2 \tag{20}$$

### A.4 Final Training Objective

The final objective combines the ELBO maximization with the VQ and orthogonality constraints. Since we frame this as maximizing the log-likelihood, the penalty terms are subtracted:

$$\begin{aligned}\mathcal{J}_{\text{total}} = & \underbrace{-\mathbb{E}_{\mathbf{s}\sim q_{\phi_s}(\mathbf{s}|\mathbf{x})}\big[\log p_\theta(\mathbf{x} \mid \mathbf{s}, \mathbf{z}_q)\big]}_{\text{Reconstruction (NLL)}} \\ & + \underbrace{\lambda_s D_{\text{KL}}(q_{\phi_s}(\mathbf{s} \mid \mathbf{x}) \,\|\, p(\mathbf{s}))}_{\text{State Regularization}} \\ & + \underbrace{\|\text{sg}[f_{\phi_z}(\mathbf{x})] - \mathbf{z}_q\|_2^2}_{\text{Codebook Loss}} + \underbrace{\beta \, \|f_{\phi_z}(\mathbf{x}) - \text{sg}[\mathbf{z}_q]\|_2^2}_{\text{Commitment Loss}} \\ & + \underbrace{\lambda_{\text{orth}} \left( \frac{\mu_s(\mathbf{x}) \cdot \text{sg}[\mathbf{z}_q]}{\|\mu_s(\mathbf{x})\|_2 \, \|\text{sg}[\mathbf{z}_q]\|_2 + \epsilon} \right)^2}_{\text{Orthogonality Penalty}}.\end{aligned} \tag{21}$$

## B Appendix: Algorithms

**Algorithm 1** Mask-Guided Disentangled Training (Compact)

**Input:** Video $\mathcal{V}$, Masks $\mathcal{M}$. **Params:** $\lambda_{orth}, \gamma, \beta$. **Init:** Codebook $\mathfrak{S}, E_{id}, E_{st}, D_\theta$.
**while** not converged **do**
  *// 1. Extract Raw Embeddings for all frames*
  $\mathcal{Z}_{all} \leftarrow \{E_{id}(\text{AlphaCLIP}(x_t, m_{t,k})) \mid \forall t \in [1, T], k \in [1, K_t]\}$
  *// 2. Identity Branch (Bipartite Aggregation)*
  Set Anchors $\mathcal{A} \leftarrow \{z^{(1,k)} \in \mathcal{Z}_{all}\}$ and Candidates $\mathcal{B} \leftarrow \mathcal{Z}_{all} \setminus \mathcal{A}$
  Compute Matching $\pi : \mathcal{B} \to \mathcal{A}$ via Bipartite Graph (maximize Cosine Sim)
  **for** $k = 1 \dots K_1$ **do** ▷ Process each unique object
    $\Omega_k \leftarrow \{z^{(1,k)}\} \cup \{b \in \mathcal{B} \mid \pi(b) = z^{(1,k)}\}$ ▷ Gather matched features
    $\bar{z}^{(k)} \leftarrow \text{Mean}(\Omega_k)$;  $z^{(k)} \leftarrow e_{j^*}$ (Nearest Codebook Entry)
    Update Codebook stats $N_{j^*}, A_{j^*}$ via EMA. $\mathcal{L}^{(k)}_{commit} \leftarrow \|\bar{z}^{(k)} - \text{sg}[z^{(k)}]\|_2^2$
  **end for**
  *// 3. State Encoding & Reconstruction*
  Sample sub-sequence $\tau \subset \{1, \dots, T\}$
  **for** $t \in \tau$ **do**
    $s^{(t,k)} \leftarrow E_{st}(\text{Pool}(\text{SuperpixelGraph}(x_t), k))$ for all $k$
    $\mathcal{L}^{(t)}_{orth} \leftarrow \sum_k (\text{CosSim}(\text{sg}[z^{(k)}], s^{(t,k)}))^2$
    $\hat{x}_t \leftarrow D_\theta(\{z^{(k)}, s^{(t,k)}\}_k)$;  $\mathcal{L}^{(t)}_{rec} \leftarrow \|x_t - \hat{x}_t\|_2^2$
  **end for**
  *// 4. Optimization*
  $\mathcal{L} \leftarrow \sum_{t \in \tau} (\mathcal{L}^{(t)}_{rec} + \lambda_{orth} \mathcal{L}^{(t)}_{orth}) + \beta \sum_k \mathcal{L}^{(k)}_{commit} + \lambda_s \text{KL}_{loss}$
  Backprop $\nabla \mathcal{L}$; Update Codebook $\mathfrak{S} \leftarrow A/N$
**end while**

**Algorithm 2** Bipartite Graph Matching for Temporal Aggregation

**Require:** Anchor Set $\mathcal{A} = \{z_{1,1}, \dots, z_{1,N}\} \in \mathbb{R}^{N \times D}$ (Frame 1 embeddings)
**Require:** Candidate Set $\mathcal{B} = \{z_{2,1}, \dots, z_{T,M}\} \in \mathbb{R}^{M \times D}$ (Embeddings from frames $2 \dots T$)
**Ensure:** Time-averaged identity embeddings $\bar{\mathcal{Z}} = \{\bar{z}_e^{(1)}, \dots, \bar{z}_e^{(N)}\}$
**Step 1: Normalization**
$\hat{\mathcal{A}} \leftarrow \mathcal{A}/\|\mathcal{A}\|_2$ ▷ L2 normalize anchors
$\hat{\mathcal{B}} \leftarrow \mathcal{B}/\|\mathcal{B}\|_2$ ▷ L2 normalize candidates
**Step 2: Construct Bipartite Graph**
$W \leftarrow \hat{\mathcal{B}} \cdot (\hat{\mathcal{A}})^\top$ ▷ Compute Cosine Similarity Matrix $W \in \mathbb{R}^{M \times N}$
**Step 3: Assignment and Aggregation**
Initialize accumulators $S \leftarrow \mathcal{A}$ ▷ Initialize sums with Frame 1 values
Initialize counters $C \leftarrow \mathbf{1}_N$ ▷ Initialize counts to 1
**for** $j = 1$ **to** $M$ **do** ▷ Iterate through all candidate objects
  $k \leftarrow \arg\max_{i \in \{1, \dots, N\}} W_{j,i}$ ▷ Find best matching anchor index
  $S_k \leftarrow S_k + \mathcal{B}_j$ ▷ Add candidate embedding to matched anchor
  $C_k \leftarrow C_k + 1$ ▷ Increment match count
**end for**
**Step 4: Final Averaging**
$\bar{\mathcal{Z}} \leftarrow S/C$ ▷ Element-wise division
**return** $\bar{\mathcal{Z}}$

# C Datasets

Our experiments involve (i) *video datasets* used for learning temporally consistent object representations and (ii) *image-centric benchmarks* used to evaluate object discovery, reconstruction quality, and downstream reasoning (Sec. 4). We summarize the datasets below and specify the preprocessing used in our setup.

## C.1 Video Training Datasets

Because PSC is trained on videos (Sec. 3), the training data must contain temporally coherent object motion and appearance variation so that temporal aggregation and identity stabilization are meaningful. We use widely adopted video datasets from the object-centric learning and compositional video reasoning literature: **MOVi-C**, **MOVi-D**, **MOVi-E**, and **CATER**. These datasets expose the model to progressively more difficult settings involving clutter, occlusion, and object interaction.

### C.1.1 MOVi-C

MOVi-C is part of the MOVi benchmark suite for multi-object video understanding. It provides synthetic videos containing multiple moving objects in controlled 3D scenes, making it well suited for learning object-centric representations with stable temporal correspondences. MOVi-C serves as a relatively controlled video source in which identity stability can be learned under moderate scene complexity.

### C.1.2 MOVi-D

MOVi-D extends the same benchmark family with more challenging scene configurations, including increased object diversity and stronger variation in appearance and spatial arrangement. Compared to MOVi-C, it introduces a harder identity-binding regime due to increased clutter and visual variability.

### C.1.3 MOVi-E

MOVi-E is the most challenging MOVi variant used here, featuring denser scenes with more severe occlusions, object interactions, and viewpoint changes. This dataset is particularly relevant for training and stress-testing temporal identity mechanisms, since objects can be partially visible, temporarily occluded, and subject to stronger appearance changes over time.

### C.1.4 CATER

CATER is a video benchmark designed for compositional spatio-temporal reasoning over moving objects and structured interactions (e.g., movement, containment, and occlusion). We include CATER because it complements the MOVi datasets with longer-range temporal dependencies and more structured object events, which are important for evaluating whether identity remains stable through nontrivial temporal transformations.

**Video preprocessing.** For each video dataset, we sample fixed-length clips and resize frames to the input resolution used by the model. Pixel intensities are normalized to $[0, 1]$. When the dataset does not provide object proposals suitable for our pipeline, we generate object masks using a class-agnostic segmentation model (Sec. 3), and use these proposals to form object-focused inputs for the identity branch.

## C.2 Object-Discovery Benchmarks

We evaluate object discovery and reconstruction on CLEVR Johnson et al. (2017), Tetrominoes Kabra et al. (2019), Objects-Room Kabra et al. (2019), Bitmoji Mozafari (2020), FFHQ Karras et al. (2020), and COCO Lin et al. (2014). These datasets are primarily image-centric and are used to measure grouping quality (ARI) and reconstruction quality (MSE/FID) as reported in Sec. 4.

#### C.2.1 CLEVR Johnson et al. (2017)

CLEVR is a synthetic benchmark with controlled object composition and clean instance-level structure. In our setting, we do not use the question-answer component; instead, we leverage CLEVR for object discovery and reconstruction. The dataset provides 70,000 training images, 15,000 validation images, and 15,000 test images. We resize all images to $64 \times 64$ and normalize pixel intensities to $[0, 1]$.

#### C.2.2 Tetrominoes Kabra et al. (2019)

Tetrominoes is a dataset of Tetris-like shapes that serves as a controlled benchmark for non-overlapping object discovery. We use approximately 100,000 images for training and 10,000 images each for validation and testing. Images are resized to $32 \times 32$ and normalized to $[0, 1]$. Each sample contains three tetromino objects drawn from a fixed set of shapes/orientations and colors.

#### C.2.3 Objects-Room Kabra et al. (2019)

Objects-Room is derived from the 3DShapes dataset Kim & Mnih (2018) and generated using the MuJoCo environment Eslami et al. (2018). It provides more cluttered synthetic scenes than CLEVR and Tetrominoes, with multiple objects placed inside a room. In our configuration, 4–6 objects appear per image, with object colors sampled at random. We use roughly 100,000 training images, 10,000 validation images, and 10,000 test images. All images are resized to $64 \times 64$ and normalized to $[0, 1]$.

#### C.2.4 Bitmoji Mozafari (2020)

Bitmoji is a dataset of cartoon face images and serves as a stylized real-domain benchmark for reconstruction quality. We resize all images to $64 \times 64$ and normalize pixel values to $[0, 1]$. The dataset contains 4,084 images in total, split into 3,500 training images and 292 images each for validation and testing.

#### C.2.5 FFHQ Karras et al. (2020)

FFHQ contains high-quality real-world face images and provides a substantially more realistic distribution than synthetic benchmarks. We use FFHQ to evaluate reconstruction quality on real images and follow the same preprocessing protocol as the corresponding baselines (resolution and normalization matched to the model configuration used in Sec. 4).

#### C.2.6 COCO Lin et al. (2014)

COCO is a large-scale dataset with complex everyday scenes containing many object instances and substantial clutter. It is used to evaluate object discovery under realistic scene complexity. For our scene decomposition analysis, we resize images to $224 \times 224$. COCO includes instance-level annotations across many object categories, which we use for ARI evaluation when ground-truth instance masks are available, following the evaluation protocol described in Sec. 4.

### C.3 Reasoning Benchmark

We evaluate downstream transfer and rationale quality using FloatingMNIST-2 (FMNIST2), which is derived from MNIST Deng (2012). FMNIST2 provides multi-object scenes with a known compositional rule, enabling controlled evaluation of transfer across tasks.

#### C.3.1 FloatingMNIST-2 (FMNIST2)

FMNIST2 is constructed by placing two randomly sampled MNIST digits on a $64 \times 64$ canvas. We consider two task variants:

- **FMNIST2-Add**: the target is the sum of the two digits;
- **FMNIST2-Sub**: the target is the absolute difference between the two digits.

We use 60,000/10,000/10,000 images for train/validation/test splits for each variant. In Sec. 4, we evaluate both in-domain performance and few-shot transfer ($k$=100) across Add↔Sub.

Table 8: Maximum number of objects (including background) for controlled datasets used in the main experiments.

| Dataset | Max. Objects |
|---|---|
| CLEVR | 11 |
| Tetrominoes | 4 |
| ObjectsRoom | 7 |
| FMNIST2 | 3 |

# D Metrics

This section defines the evaluation metrics used in Sec. 4. We report metrics for object discovery, reconstruction quality, visual reasoning, video-centric identity stability, and representation grounding.

## D.1 Foreground-Adjusted Rand Index (ARI)

For object discovery, we evaluate instance grouping quality using the foreground-adjusted Adjusted Rand Index (ARI), following standard practice in object-centric learning. ARI compares the predicted object assignment of each foreground pixel against the ground-truth instance assignment, while ignoring background pixels.

Let $\Omega_{\text{fg}}$ denote the set of foreground pixels, and let $y_p$ and $\hat{y}_p$ denote the ground-truth and predicted instance labels for pixel $p \in \Omega_{\text{fg}}$. Let $n_{ij}$ be the number of foreground pixels assigned to ground-truth object $i$ and predicted object $j$, with row and column sums $a_i = \sum_j n_{ij}$ and $b_j = \sum_i n_{ij}$, and let $n = \sum_{ij} n_{ij} = |\Omega_{\text{fg}}|$. Then ARI is defined as

$$\text{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \frac{\left(\sum_i \binom{a_i}{2}\right)\left(\sum_j \binom{b_j}{2}\right)}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\left(\sum_i \binom{a_i}{2}\right)\left(\sum_j \binom{b_j}{2}\right)}{\binom{n}{2}}}. \tag{22}$$

Higher values indicate better agreement between predicted and ground-truth object assignments.

## D.2 Mean Squared Error (MSE)

To assess reconstruction fidelity, we compute the pixel-wise mean squared error between each input image and its reconstruction. Let $x^i$ denote the $i$-th input image and $\hat{x}^i$ its reconstruction. Over $N$ images, MSE is defined as

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} \left\| x^i - \hat{x}^i \right\|_2^2 . \tag{23}$$

Lower values indicate more accurate reconstructions.

## D.3 Fréchet Inception Distance (FID)

We use Fréchet Inception Distance (FID) to measure the perceptual similarity between reconstructed and real images. FID compares the first- and second-order statistics of features extracted by a pretrained Inception network.

Let $\phi(\cdot)$ denote the Inception feature extractor, and let $(\mu_r, \Sigma_r)$ and $(\mu_g, \Sigma_g)$ denote the empirical mean and covariance of features from the real and reconstructed image sets, respectively. Then FID is defined as

$$\text{FID}(X_r, X_g) = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\Big(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}\Big) , \tag{24}$$

where $X_r$ and $X_g$ denote the real and reconstructed image sets. Lower values indicate that reconstructed images are closer to the real-image distribution.

## D.4 Accuracy

For reasoning tasks, we report classification accuracy, defined as the fraction of correctly predicted labels. Let $y^i$ and $\hat{y}^i$ denote the ground-truth and predicted labels for the $i$-th example. Over $N$ examples,

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\big[\hat{y}^i = y^i\big] , \tag{25}$$

where $\mathbb{I}[\cdot]$ is the indicator function. Higher values indicate better predictive performance.

### D.5 F1 Score

We also report F1 score for reasoning outputs, which balances precision and recall. Let precision and recall be defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{26}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. The F1 score is then

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{27}$$

Higher values indicate better agreement with the target labels or rationale annotations. For multi-class settings, we compute the standard class-averaged F1 used by the corresponding benchmark.

### D.6 Hungarian Matching Coefficient (HMC)

For reasoning-oriented objectives, we evaluate rationale alignment by matching predicted slot-level property vectors to ground-truth object property vectors. Because object slots are permutation-invariant, we first solve a bipartite matching problem.

Let $\{\hat{r}_k\}_{k=1}^{K}$ denote the predicted slot-level rationale vectors and $\{r_j\}_{j=1}^{K}$ denote the corresponding ground-truth object-level rationale vectors. We define the pairwise matching cost as

$$C_{kj} = \|\hat{r}_k - r_j\|_2^2 . \tag{28}$$

We then compute the optimal permutation

$$\pi^\star = \arg\min_{\pi \in S_K} \sum_{k=1}^{K} C_{k,\pi(k)}, \tag{29}$$

where $S_K$ is the set of all permutations of $\{1, \ldots, K\}$. The Hungarian Matching Coefficient is the average matched cost:

$$\text{HMC} = \frac{1}{K} \sum_{k=1}^{K} C_{k,\pi^\star(k)}. \tag{30}$$

Lower values indicate better alignment between predicted and ground-truth rationales.

### D.7 Re-identification Accuracy

To directly evaluate temporal identity stability, we measure whether the same physical object is assigned the same predicted identity across frames. Let there be $M$ ground-truth object tracks, and let $\mathcal{T}_m$ denote the set of visible frames for track $m$. Let $\hat{c}_{t,m}$ denote the predicted identity code (or matched identity label) for object $m$ at frame $t$, and let $t_m^{(0)}$ be the first visible frame of that track. We define re-identification accuracy as

$$\text{ReID} = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{|\mathcal{T}_m| - 1} \sum_{t \in \mathcal{T}_m \setminus \{t_m^{(0)}\}} \mathbb{I}\Big[\hat{c}_{t,m} = \hat{c}_{t_m^{(0)},m}\Big] . \tag{31}$$

Higher values indicate more stable identity assignments over time.

### D.8 Identity-Switch Rate

Identity-switch rate measures how often the predicted identity of an object changes along its track. Using the same notation as above, we define

$$\text{IDSwitch} = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{|\mathcal{T}_m| - 1} \sum_{t \in \mathcal{T}_m \setminus \{t_m^{(0)}\}} \mathbb{I}\big[\hat{c}_{t,m} \neq \hat{c}_{t^-,m}\big] , \tag{32}$$

where $t^-$ denotes the previous visible frame of object $m$ before frame $t$. Lower values indicate better temporal consistency.

### D.9 Object Retrieval Accuracy

To evaluate whether the learned representation is grounded in the target object instance, we measure cross-frame object retrieval. For each query object $m$ in frame $t$, let $z_{t,m}$ denote its representation, and let $\{z_{t',j}\}_{j=1}^{J}$ denote the candidate object representations in another frame $t'$, one of which corresponds to the same physical object. We retrieve the nearest candidate under cosine similarity:

$$\hat{j} = \arg\max_{j \in \{1,\dots,J\}} \frac{z_{t,m}^{\top} z_{t',j}}{\|z_{t,m}\|_2 \, \|z_{t',j}\|_2}. \tag{33}$$

Let $j^{\star}$ denote the ground-truth matching object in frame $t'$. Object retrieval accuracy is then

$$\mathrm{ObjRet} = \frac{1}{Q} \sum_{q=1}^{Q} \mathbb{I}\left[\hat{j}^{(q)} = j^{\star(q)}\right], \tag{34}$$

where $Q$ is the total number of query-target pairs. Higher values indicate that the representation preserves object-specific identity across frames.

### D.10 Background Invariance

A grounded object representation should remain stable when the target object is preserved but the surrounding context is perturbed. Let $z_m$ denote the representation of object $m$ under the original input, and let $\tilde{z}_m^{\mathrm{bg}}$ denote the representation obtained after replacing or perturbing the background while keeping the target object unchanged. We define background invariance as the mean cosine similarity

$$\mathrm{BgInv} = \frac{1}{N} \sum_{i=1}^{N} \frac{z_i^{\top} \tilde{z}_i^{\mathrm{bg}}}{\|z_i\|_2 \, \|\tilde{z}_i^{\mathrm{bg}}\|_2}. \tag{35}$$

Higher values indicate that the representation is insensitive to irrelevant contextual changes and remains anchored to the object itself.

### D.11 Object-Removal Sensitivity

To test whether the representation depends on the presence of the target object rather than only on scene context, we remove the target object region and recompute the representation. Let $z_m$ denote the original representation of object $m$, and let $\tilde{z}_m^{\mathrm{rm}}$ denote the representation obtained after masking or deleting the target object while leaving the rest of the scene unchanged. We define object-removal sensitivity as

$$\mathrm{ObjRem} = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{z_i^{\top} \tilde{z}_i^{\mathrm{rm}}}{\|z_i\|_2 \, \|\tilde{z}_i^{\mathrm{rm}}\|_2}\right). \tag{36}$$

Higher values indicate that the representation changes substantially when object evidence is removed, which is consistent with stronger grounding in the target instance.

Table 9: Training throughput (iterations/second) for the object-discovery task. All measurements should be collected under the same hardware and batch-size setting.

| Method | CLEVR | Tetrominoes | Objects-Room | Bitmoji | FFHQ | COCO |
|---|---|---|---|---|---|---|
| SA (Locatello et al., 2020) | 1.06 | 8.42 | 4.18 | 2.34 | 1.91 | – |
| Block-Slot (Singh et al., 2022) | 0.72 | 5.61 | 2.76 | 1.54 | 1.28 | – |
| IMPLICIT (Chang et al., 2022) | 0.83 | 6.17 | 3.09 | 1.69 | 1.41 | 0.92 |
| CoSA (Kori et al., 2023) | 0.57 | 4.08 | 2.01 | 1.18 | 0.97 | 0.69 |
| AdaSlot (Fan et al., 2024) | 0.95 | 7.31 | 3.66 | 2.07 | 1.69 | 1.08 |
| SPOT (Kakogeorgiou et al., 2024) | 0.69 | 5.24 | 2.58 | 1.73 | 1.44 | 0.85 |
| PSC | 0.88 | 6.86 | 3.34 | 1.96 | 1.62 | 1.01 |

## E Training and Computational Details

We train all models using the Adam optimizer with a learning rate of $4 \times 10^{-4}$ and a batch size of 16. Unless otherwise noted, training uses early stopping with a patience of 5 and runs for at most $\min(40 \text{ epochs}, 40{,}000 \text{ steps})$. We apply a linear learning-rate warmup for the first 10,000 steps, followed by a Reduce-on-Plateau scheduler with decay factor 0.5. This configuration is used for the main PSC model as well as the comparison models, unless the original implementation of a baseline requires a different scheduler for stable reproduction.

For PSC, training is performed on fixed-length video clips, since the identity branch aggregates object-level representations across time. During training, input frames are resized to the task-specific resolution described in Appendix C, normalized to $[0, 1]$, and converted into object-focused inputs using the mask-generation pipeline described in Sec. 3. Unless otherwise noted, we use the same optimization hyperparameters for the object-discovery experiments and the downstream FMNIST2 transfer experiments, with task-specific heads trained on top of the learned object representations.

We run all experiments on a compute cluster equipped with NVIDIA Tesla T4 16GB GPUs and Intel(R) Xeon(R) Gold 6230 CPUs. Throughput is reported in iterations per second (it/s). These measurements may vary slightly across systems depending on I/O load, background processes, and implementation details, so they should be interpreted as approximate efficiency indicators rather than absolute hardware-independent benchmarks.

To support reproducibility, we will release the experimental configuration files, training scripts, inference scripts, and dataset-generation scripts used in our experiments.

**Implementation notes.** For fair comparison, we use the official implementations of the baselines whenever available and preserve their default architectural components unless changes are required to match input resolution or evaluation protocol. When a baseline is adapted to our setting, we report this explicitly in the corresponding experiment section. For throughput reporting, all models are evaluated under the same batch size and hardware configuration.

Table 10: Training throughput (iterations/second) for the FMNIST2 reasoning task. Results should be measured under the same few-shot transfer setup and hardware configuration.

| Method | FMNIST2 |
|---|---|
| CNN | 14.20 |
| SA (Locatello et al., 2020) | 8.06 |
| Block-Slot (Singh et al., 2022) | 5.71 |
| CoSA (Kori et al., 2023) | 3.92 |
| AdaSlot (Fan et al., 2024) | 7.11 |
| SPOT (Kakogeorgiou et al., 2024) | 5.18 |
| PSC | 5.12 |

Table 11: Video pretraining throughput (iterations/second) for the PSC representation-learning stage. This table is useful when reporting the cost of temporal training separately from downstream evaluation.

| Video Dataset | PSC (it/s) |
|---|---|
| MOVi-C | 1.42 |
| MOVi-D | 1.18 |
| MOVi-E | 0.94 |
| CATER | 1.07 |