Selective Synergistic Learning for Video Object-Centric Learning

WonJun Moon[1†] and Jae-Pil Heo[2⋆]

[1] KAIST, South Korea
[2] Sungkyunkwan University, South Korea

Abstract. Typical video object-centric learning (VOCL) approaches employ slot-based frameworks that rely on reconstruction-driven encoder–decoder architectures, where learning is mediated by two spatial maps: attention maps from the encoder and object maps from the decoder. As these two distinct maps exhibit different properties, a recent dense alignment strategy attempted to reconcile this discrepancy by enforcing agreement across all spatio-temporal patches via contrastive learning. However, this indiscriminate alignment inadvertently propagates the inherent weaknesses of each module, such as noisy encoder predictions and blurred decoder boundaries. Moreover, computing dense similarities across all pairs incurs a computational cost quadratic in the total number of spatio-temporal patches, severely limiting scalability. Motivated by this, we propose Selective Synergistic Learning (SSync). Instead of exhaustive patch-to-patch alignment, SSync prevents error propagation by selectively distilling only the most reliable cues: leveraging the encoder strictly for boundary refinement and the decoder for interior denoising. This is realized via a pseudo-labeling with linear complexity, eliminating the need for quadratic spatial comparisons. Also, to prevent the reinforcement of architectural biases like slot redundancy, we introduce a transitive pseudo-label merging that consolidates overlapping slots based on spatio-temporal activation consistency. Extensive studies demonstrate that SSync improves decomposition quality and serves as a versatile, plug-and-play module while also exhibiting exceptional robustness to slot configurations. Code is available at github.com/wjun0830/SSync.

Keywords: Video Object-Centric Learning · Object Discovery · Object Representation Learning · Self-Supervised Representation Learning

1 Introduction

Object-centric learning is a fundamental paradigm for structured visual reasoning, enabling downstream applications such as object editing, generation, and scene understanding [1–7]. By decomposing scenes into object representations, object-centric frameworks provide an interpretable and modular representation space

† Work done at Sungkyunkwan University.
⋆ Corresponding author

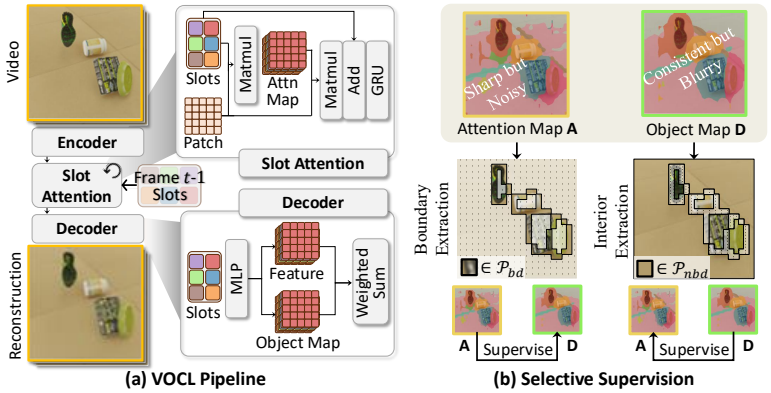

Fig. 1: Overall flow and motivation. (a) Video frames are sequentially processed, where slots are recurrently updated based on the previous frame's slots. (b) Slot attention map captures sharp boundaries but contains noise, while the decoder object map offers a consistent representation but suffers from blurry edges. To leverage their complementary strengths, SSync reciprocally distills the sharp boundary cues from the attention map and the consistent semantics of the decoder object map.

that facilitates reasoning beyond pixel-level perception [8]. Among them, slot attention [9] has emerged as a dominant method, grouping patch features into latent slots and reconstructing the input through a decoder that implicitly produces object maps. It has recently been extended to videos, leveraging temporal cues to achieve consistent object discovery and tracking across frames [10–15].

Despite their empirical success, slot architectures suffer from a fundamental structural mismatch. Learning is mediated by two spatial maps: the encoder's attention maps and the decoder's reconstructed object maps. Ideally, these maps should be logically consistent, as both branches are jointly optimized through a shared reconstruction objective. However, in practice, they exhibit distinct inductive biases. The encoder, often built upon high-capacity vision backbones [16], produces spatially sharp yet noisy assignments, exhibiting high-frequency sensitivity. In contrast, the decoder, typically implemented as a lightweight MLP, imposes a low-frequency spatial prior, yielding smooth and temporally coherent but blurry object boundaries [14]. This structural asymmetry induces persistent misalignment between patch assignments and rendering regions, resulting in unstable slot semantics and fragmented object decomposition.

A straightforward solution is to enforce consistency between these two spatial maps. Recently, SRL [14] attempts to bridge this gap through dense contrastive alignment across all spatio-temporal patches. While conceptually appealing, we argue that dense alignment is fundamentally misaligned with the heterogeneous inductive biases of the encoder and decoder since it treats every patch as an equally reliable teacher. Consequently, it inadvertently propagates the inherent failure modes of each branch. Also, computing similarities across all spatio-temporal patches incurs a quadratic memory cost, severely limiting its applicability to long sequences or high-resolution videos.

In this work, we introduce a selective alignment principle: mutual supervision should occur only where each branch provides reliable cues. A key insight is that, once reliable regions are correctly identified, the alignment mechanism itself can be remarkably simple. We introduce Selective Synergistic Learning (SSync), a selective mutual distillation framework that filters supervision at the patch level instead of enforcing global agreement. Concretely, we exploit the encoder's boundary sensitivity to refine decoder masks at object boundaries, while leveraging the decoder's spatial coherence to denoise encoder assignments within interior regions. Specifically, we exploit local consistency cues to discern object boundaries within the encoder and coherent interior regions within the decoder, and formulate an efficient supervisory signal through pseudo-labeling. By aligning only in these reliable regions, SSync mitigates error propagation inherent in dense objectives while substantially reducing memory complexity. Our contributions are: (1) **Reliability-based selective distillation.** We identify the spatial complementarity between encoder and decoder expertise and reformulate mutual learning as a selective cross-distillation framework, (2) **Redundancy control via transitive merging.** To stabilize on-the-fly pseudo-labeling during optimization, we introduce a transitive merging strategy, and (3) **Strong performance & practical generalization.** SSync achieves state-of-the-art results across VOCL benchmarks, improves memory efficiency relative to dense alignment, and remains robust to varying slot configurations.

2 Related Work

2.1 Object-Centric Representation Learning

Object-centric learning seeks to decompose scenes into discrete structural entities [17–19]. Slot Attention [9] established a foundational framework by iteratively grouping spatial features into latent slots and reconstructing the image via a shared decoder. This encoder–decoder formulation enables unsupervised object discovery while maintaining permutation invariance over slots.

Recent works have extended slot-based learning to dynamic visual scenes [15, 20]. SAVi [12] and SAVi++ [13] leveraged motion cues such as optical flow and depth to promote temporal consistency. STEVE [11] incorporated transformer-based models to capture complex object interactions. Videosaur [10] introduced objectives that encourage grouping of motion-consistent regions, and SlotContrast [21] enhanced slot discriminability via temporal contrastive learning. While these significantly improve object discovery in videos, they predominantly rely on a reconstruction objective, which implicitly assumes a convergence between encoder attention and decoder object maps. In practice, however, these two representations exhibit distinct inductive biases, leading to spatial discrepancies.

To address this discrepancy, SRL [14] introduced a mutual refinement process via a dense alignment strategy, employing a contrastive objective across all spatio-temporal patches. While this reduces inconsistency, it overlooks the unique strengths and inherent limitations of each branch by uniformly enforcing agreement across all regions. This indiscriminate alignment inevitably conflates

complementary strengths with architectural weaknesses, all while incurring prohibitive memory costs. In contrast, we explicitly identify the expertise of each branch and instantiate synergy through selective cross-distillation, thereby preventing error propagation and ensuring linear scalability.

2.2 Pseudo-Labeling

Pseudo-labeling has been widely adopted in semi-supervised and self-supervised learning to exploit high-confidence predictions as supervisory signals [22–27]. Self-training approaches such as Noisy Student [28] and Meta Pseudo Labels [29] also demonstrate that iterative refinement of pseudo-labels can significantly improve representation quality.

However, directly applying pseudo-labeling to slot learning introduces unique challenges. Since pseudo-labels are generated on-the-fly from the model's own predictions, they are inherently unstable; early training errors can easily propagate via pseudo-labeling, making the framework highly vulnerable to noise. To resolve this, we introduce a transitive pseudo-label merging that analyzes spatio-temporal activation overlap to detect redundancy and consolidate slot identities globally. This redundancy-aware refinement stabilizes selective supervision and prevents feedback loops caused by imperfect on-the-fly pseudo-labels.

3 Selective Synergistic Learning

3.1 Preliminaries

Video Slot Learning and Spatial Maps. We consider an input video represented as a grid of patch tokens. Let $z_{t,p}$ denote the feature of patch $p \in \{1, \dots, P\}$ at time $t \in \{1, \dots, T\}$, where $P = H \times W$. A slot attention encoder predicts S slot prototypes. By computing the dot product between the projected patch queries $\mathbf{q}_{t,p}$ (derived from $z_{t,p}$) and the slot keys $\mathbf{k}_{t,s}$, the encoder generates an encoder attention map over slots for each patch:

$$\mathbf{A}_{t,p} = \mathrm{Softmax}\big(\mathbf{q}_{t,p}^{\top}\mathbf{k}_{t,1:S}\big), \tag{1}$$

where $\mathbf{A}_{t,p,s}$ is the probability that patch $z_{t,p}$ is explained by slot s. A decoder reconstructs the input and produces a decoder object map $\mathbf{D} \in \mathbb{R}^{T\times P\times S}$, which is normalized over the slot dimension such that $\sum_{s=1}^{S} \mathbf{D}_{t,p,s} = 1$ for all t and p. Both maps provide patch-to-slot assignments but are generated by different modules and thus exhibit different error characteristics.

Bias Misalignment in Encoder-Decoder Maps. Ideally, the attention map $\mathbf{A}$ and the object map $\mathbf{D}$ should be consistent, since they are optimized jointly under reconstruction-driven learning. In practice, however, $\mathbf{A}$ and $\mathbf{D}$ are persistently misaligned since they inherit distinct spatial characteristics from different modules. The encoder typically produces spatially sharp attention maps but is vulnerable to noisy assignments, while the decoder tends to yield spatio-temporally coherent maps but with blurred object boundaries. This heterogeneity

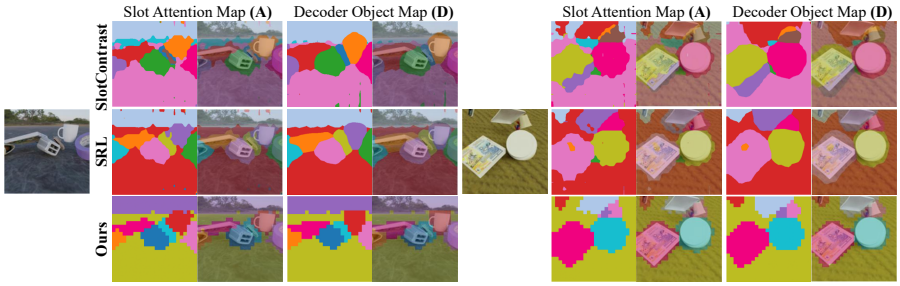

Fig. 2: Visualization of the attention map A and object map D. Compared to SlotContrast [21], SRL [14] reduces noise and produces sharper slot assignments. However, dense alignment also propagates both noise and blur across the two branches. Consequently, noisy and spatially blurry representations are still observed in both maps, indicating incomplete resolution of encoder–decoder discrepancy.

implies that reliability is region-dependent: boundary regions benefit from the encoder's sharpness, whereas interior regions benefit from the decoder's coherence.

Limitations of Dense Alignment. SRL [14] attempted to reduce discrepancy between **A** and **D** via dense alignment (e.g., contrastive objectives [30, 31] across the full spatio-temporal volume). While straightforward, it implicitly assumes that all patches provide equally reliable supervision; when this assumption fails, indiscriminate agreement propagates erroneous signals from unreliable artifacts. Consequently, SRL only partially resolves the discrepancy, as each module inadvertently reinforces the other's inherent flaws. We provide quantitative and qualitative evidence of this failure in Fig. 2 and Tab. 1. Also, dense patch-wise objective requires pairwise spatio-temporal comparisons, resulting in quadratic complexity $\mathcal{O}((T \cdot H \cdot W)^2)$, which limits its scalability to long or high-resolution videos.

Table 1: Encoder-decoder consistency. We used symmetric Adjusted Rand Index (ARI) and asymmetric mean Best Overlap (mBO).

Method	ARI $\mathbf{A} \leftrightarrow \mathbf{D}$	mBO $\mathbf{A} \rightarrow \mathbf{D}$	mBO $\mathbf{D} \rightarrow \mathbf{A}$
SlotContrast	73.0	63.6	62.9
SRL	77.6	65.4	60.8
Ours	**85.7**	**72.2**	**68.6**

Motivation: Selective Alignment Principle. Above observations reveal that dense alignment requires simultaneous reliability of both maps, while encoder-decoder asymmetry during optimization often invalidates this assumption. Therefore, we propose a selective alignment principle: mutual supervision should be applied only where each branch is structurally reliable. Concretely, we leverage encoder-driven boundaries and decoder-driven interiors to align the counterpart branch, proving that minimal supervision is sufficient for robust alignment once reliable regions are identified. This prevents error propagation caused by indiscriminate agreement and eliminates the need for dense spatio-temporal pairwise comparisons.

3.2 Structuring Pseudo-Labels

To instantiate the selective alignment principle, we first construct structured pseudo-labels from the model's internal spatial maps. These pseudo-labels serve as teacher signals for cross-module refinement. For each spatio-temporal patch at

(t, p), we derive hard slot assignments from the probabilistic encoder attention map $\mathbf{A}$ and decoder object map $\mathbf{D}$:

$$\hat{s}^{A}_{t,p} = \arg\max_{s} \mathbf{A}_{t,p,s}, \qquad \hat{s}^{D}_{t,p} = \arg\max_{s} \mathbf{D}_{t,p,s}. \tag{2}$$

These provide candidate targets, whose reliability varies across regions.

Local Consistency Analysis. To identify structurally reliable regions, we measure local slot consistency separately on the encoder attention assignments and the decoder object-map assignments. Let $\mathcal{N}_{\text{sp}}(p)$ denote the spatial 8-neighborhood of patch p on the $H \times W$ patch grid. We define the spatiotemporal neighborhood of a patch (t, p) by augmenting $\mathcal{N}_{\text{sp}}(p)$ with temporally adjacent patches at the same spatial location to further detect motion edges:

$$\mathcal{N}(t, p) = \{(t, q) \mid q \in \mathcal{N}_{\text{sp}}(p)\} \cup \{(t-1, p), (t+1, p)\}, \tag{3}$$

where out-of-range temporal indices are omitted. For $\mathbf{X} \in \{\mathbf{A}, \mathbf{D}\}$, we quantify the local disagreement and agreement counts as:

$$c^{\neq,\mathbf{X}}_{(t,p)} = \sum_{(t',q)\in\mathcal{N}(t,p)} \mathbb{I}\big[\hat{s}^{\mathbf{X}}_{t',q} \neq \hat{s}^{\mathbf{X}}_{t,p}\big], \qquad c^{=,\mathbf{X}}_{(t,p)} = \sum_{(t',q)\in\mathcal{N}(t,p)} \mathbb{I}\big[\hat{s}^{\mathbf{X}}_{t',q} = \hat{s}^{\mathbf{X}}_{t,p}\big]. \tag{4}$$

Boundary Region Selection. Boundary regions are characterized by local disagreement while still maintaining semantic grounding. Since encoder attention assignments are typically sharper, we detect boundary candidates from $\mathbf{A}$:

$$\mathcal{P}_{\text{bd}} = \left\{(t, p) \mid c^{\neq,\mathbf{A}}_{(t,p)} \geq n_{\text{bd}} \ \wedge \ c^{=,\mathbf{A}}_{(t,p)} \geq 1\right\}, \tag{5}$$

where n_{bd} is a predefined threshold that controls the sensitivity of boundary detection. This criterion captures meaningful object transitions while excluding isolated noisy assignments by requiring at least one additional agreeing neighbor.

Interior Region Selection. In contrast, interior (non-boundary) regions exhibit high local consistency, which is more reliably reflected in the decoder object map assignments. We thus define the set of interior patches $\mathcal{P}_{\text{nbd}}$ using $\mathbf{D}$:

$$\mathcal{P}_{\text{nbd}} = \left\{(t, p) \mid c^{\neq,\mathbf{D}}_{(t,p)} < n_{\text{nbd}}\right\}, \tag{6}$$

where n_{nbd} controls the strictness of interior selection; a patch is included in $\mathcal{P}_{\text{nbd}}$ only if the number of differing neighbor assignments is strictly smaller than n_{nbd}. Consequently, these boundary and interior sets ($\mathcal{P}_{\text{bd}}$ and $\mathcal{P}_{\text{nbd}}$) provide reliable regions for asymmetric cross-distillation: boundary patches supervise the decoder using encoder-derived pseudo-labels, while interior patches supervise the encoder using decoder-derived pseudo-labels.

From a broader perspective, our selection acts as a relaxed morphological erosion [32] parameterized by a 3×3 kernel of all ones. Unlike erosion, which strictly requires all neighboring pixels within a kernel to agree (making it highly susceptible to noise patches), our parameterized spatiotemporal formulation explicitly filters noise and overcomes standard erosion's lack of temporal awareness.

3.3 Transitive Pseudo-Label Merging

Although selective alignment restricts supervision to structurally reliable regions of each module, we note that pseudo-labels are still imperfect as they are generated on-the-fly from intermediate model predictions. As a result, they may inherit systematic semantic errors such as over-fragmentation[3], where a single object is split across multiple slot identities. When such fragmented assignments are directly used as supervision targets, the labels themselves become inherently flawed. This inadvertently reinforces over-fragmentation by forcing the model to associate a single semantic object with multiple, conflicting slot identities. To prevent this instability, we refine pseudo-labels before applying selective alignment losses. Specifically, we detect redundant slots based on spatio-temporal activation overlap and consolidate them using a transitive connectivity criterion.

For each slot s, we determine whether a spatio-temporal patch is active by thresholding its attention value against the slot-wise mean activation μ_s:

$$\mathbf{M}_{t,p,s} = \mathbb{I}\Big[\mathbf{A}_{t,p,s} > \mu_s\Big], \qquad \mu_s = \frac{1}{TP}\sum_{t,p}\mathbf{A}_{t,p,s}. \tag{7}$$

This yields a binary mask $\mathbf{M}$ that identifies regions where each slot exhibits above-average activation over space and time. Using the binary active-region masks $\mathbf{M}$, we measure the overlap between slots (s, s') via frame-averaged IoU:

$$\mathrm{IoU}(s,s') = \frac{1}{T}\sum_{t=1}^{T}\frac{\sum_p \mathbf{M}_{t,p,s}\mathbf{M}_{t,p,s'}}{\sum_p \mathbf{M}_{t,p,s} + \sum_p \mathbf{M}_{t,p,s'} - \sum_p \mathbf{M}_{t,p,s}\mathbf{M}_{t,p,s'}}. \tag{8}$$

We construct a redundancy graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $s \in \mathcal{V}$ corresponds to a slot identity. An undirected edge $(s, s') \in \mathcal{E}$ is added if

$$\mathrm{IoU}(s,s') > \tau_{\mathrm{merge}}, \tag{9}$$

indicating that the two slots exhibit substantial spatio-temporal overlap and are therefore likely to represent the same object.

Since redundancy may occur transitively (e.g., s_1 overlaps with s_2, and s_2 overlaps with s_3), we group redundant slots by computing the connected components of $\mathcal{G}$. Each connected component $\mathcal{C}_k \subset \mathcal{V}$ represents a cluster of mutually redundant slots. We denote the set of all disjoint clusters as $\mathbb{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_m\}$. For each component $\mathcal{C}_k$, we select a dominant slot identity:

$$s_k^\star = \arg\max_{s\in\mathcal{C}_k}\sum_{t,p}\mathbb{I}\big[\hat{s}^A_{t,p} = s\big], \tag{10}$$

where the slot with the largest footprint is chosen as the representative identity.

[3] Over-fragmentation arises when multiple slots redundantly cooperate to reconstruct a single object in order to minimize reconstruction loss [33].

We then define a relabeling function $\phi(\cdot)$ and apply ϕ to unify pseudo-label assignments to provide coherent object identities for selective alignment:

$$\phi(s) = \begin{cases} s_k^\star & \text{if } s \in \mathcal{C}_k, \\ s & \text{otherwise.} \end{cases} \tag{11}$$

$$\hat{s}_{t,p}^A \leftarrow \phi(\hat{s}_{t,p}^A), \qquad \hat{s}_{t,p}^D \leftarrow \phi(\hat{s}_{t,p}^D). \tag{12}$$

Using the relabeled pseudo-labels, we recompute the local consistency counts and re-derive the boundary/interior sets $\mathcal{P}_{\text{bd}}$ and $\mathcal{P}_{\text{nbd}}$ via Eq. (4)-(6). For simplicity, we reuse $\hat{s}_{t,p}^A$ and $\hat{s}_{t,p}^D$ to denote the merged pseudo-labels after applying $\phi(\cdot)$.

3.4 Training Objective

Given the refined pseudo-labels obtained after transitive merging, we perform asymmetric cross-distillation between the encoder attention map $\mathbf{A}$ and the decoder object map $\mathbf{D}$. For boundary patches at $(t,p) \in \mathcal{P}_{\text{bd}}$, the decoder is supervised using one-hot pseudo-labels derived from the encoder:

$$\mathcal{L}_{\text{bd}} = \frac{1}{|\mathcal{P}_{\text{bd}}|} \sum_{(t,p)\in\mathcal{P}_{\text{bd}}} \left\| \mathbf{D}_{t,p} - \text{onehot}(\hat{s}_{t,p}^A) \right\|_2^2, \tag{13}$$

while the interior $(t,p) \in \mathcal{P}_{\text{nbd}}$ of the encoder is supervised from the decoder as:

$$\mathcal{L}_{\text{nbd}} = \frac{1}{|\mathcal{P}_{\text{nbd}}|} \sum_{(t,p)\in\mathcal{P}_{\text{nbd}}} \left\| \mathbf{A}_{t,p} - \text{onehot}(\hat{s}_{t,p}^D) \right\|_2^2. \tag{14}$$

We adopt an MSE objective rather than cross-entropy, as its scale compatibility with the reconstruction loss enables stable joint optimization without loss reweighting, and its bounded gradients improve robustness to imperfect pseudo-labels during early training [34, 35].

Finally, the selective alignment losses are combined with the base objective $\mathcal{L}_{\text{base}}$, which includes reconstruction and temporal slot contrastive objective [21] following SRL [14]. Following a warm-up phase covering the first 30% of the total iterations η_t, the complete training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \mathbb{I}[\eta > 0.3\eta_{\text{t}}] \cdot \lambda_{\text{SSync}} \left(\mathcal{L}_{\text{bd}} + \mathcal{L}_{\text{nbd}} \right), \tag{15}$$

where η denotes the current iteration and λ_{SSync} is the coefficient for SSync.

4 Experiments

4.1 Evaluation Settings

Datasets and Metrics. Following prior works [10, 14, 21], we run experiments on three standard VOCL benchmarks: MOVi-C and MOVi-E [36], and

Table 2: Experimental results. Results are averaged across 3 runs. The best and the second best results are denoted by red and orange.

Method	Venue	MOVi-C		MOVi-E		YouTube-VIS	
		FG-ARI↑	mBO↑	FG-ARI↑	mBO↑	FG-ARI↑	mBO↑
SAVi [12]	ICLR'22	22.2	13.6	42.8	16.0	-	-
STEVE [11]	NeurIPS'22	36.1	26.5	50.6	26.6	15.0	19.1
VideoSAUR [10]	NeurIPS'23	64.8	**38.9**	73.9	**35.6**	28.9	26.3
VideoSAURv2 [10]	NeurIPS'23	–	–	77.1	34.4	31.2	29.7
SlotContrast [21]	CVPR'25	69.3	32.7	82.9	29.2	38.0	33.7
SRL [14]	ICLR'26	74.3	34.5	81.9	29.3	**42.9**	**35.6**
SlotCurri [15]	CVPR'26	**77.6**$_{\pm 0.9}$	32.8$_{\pm 0.2}$	**83.7**$_{\pm 0.2}$	28.9$_{\pm 0.7}$	**44.8**$_{\pm 1.2}$	35.5$_{\pm 2.2}$
SSync (Ours)	ECCV'26	**79.4**$_{\pm 0.6}$	**39.5**$_{\pm 0.1}$	**84.0**$_{\pm 0.9}$	**34.8**$_{\pm 1.9}$	42.6$_{\pm 0.2}$	**38.7**$_{\pm 0.6}$

YouTube-VIS (YTVIS) 2021 [37–39]. The primary challenge of MOVi-C is over-fragmentation due to complex object interactions, whereas MOVi-E contains a large number of small-scale objects, making boundary precision critical. YTVIS further evaluates robustness under real-world video dynamics. To further assess generalizability, we also evaluate under the RandSF.Q protocol [20], benchmarking on lower-resolution variants of MOVi-C and MOVi-D, as well as the High-Quality YTVIS dataset. Lastly, to verify generalization beyond video domain, we report image-level object-centric performance on MOVi-E and COCO2017 [40].

For metrics, we use Foreground ARI (FG-ARI) [41,42] and mBO [17]. FG-ARI measures permutation-invariant clustering consistency between predicted slot assignments and ground-truth (GT) masks over foreground pixels. mBO evaluates object-level coverage by computing the maximum Intersection-over-Union (IoU) between each GT object and predicted slots. We also report ARI and mIoU when available in prior work.

Implementation Details. We adopt prior training protocols [14, 20, 21] to ensure fair comparison; all architectural components and base objectives follow the respective baselines. SSync additionally introduces n_{bd}, n_{nbd}, λ_{SSync} and τ_{merge}. Among them, we fix $n_{\text{bd}} = n_{\text{nbd}} = 1$ and $\lambda_{\text{SSync}} = 1.0$ across all datasets, which demonstrates SSync's robustness and its minimal requirement for dataset-specific tuning. We only adjust the merging threshold τ_{merge}, setting it to 0.7 for MOVi-C, 0.65 for MOVi-E, and 0.6 for YTVIS. This accounts for varying object densities, reflecting theoretically established principles in clustering where overlap criteria must adapt to spatial density [43, 44]. Importantly, we observe that performance remains consistent within a reasonable range around 2/3, indicating that SSync does not rely on exhaustive threshold tuning. Details are in the Appendix.

4.2 Comparison to the State-of-the-art (SOTA) Methods

Video Benchmarks. Comparison to SOTA VOCL methods is shown in Tab. 2, where SSync consistently achieves superior or highly competitive performance compared to prior approaches across benchmarks. On MOVi-C, where objects

Table 3: Results under the RandSF.Q protocol [20]. *tsim* denotes the time similarity loss from VideoSAUR [10], and SSC refers to the temporal slot contrastive loss from SlotContrast [21]. All baselines are reproduced using their official implementations.

Method	MOVi-C				MOVi-D				HQ-YTVIS			
	ARI	FGARI	mBO	mIoU	ARI	FGARI	mBO	mIoU	ARI	FGARI	mBO	mIoU
RandSF.Q_{tsim} [20]	70.7	63.3	31.1	28.1	39.3	70.5	25.6	24.3	39.2	56.3	37.2	37.0
RandSF.Q_{tsim}+SSync	**73.0**	**67.3**	**32.3**	**29.9**	**45.5**	**71.2**	**27.6**	**25.7**	**42.9**	**58.6**	**39.4**	**39.2**
RandSF.Q_{ssc} [20]	52.7	67.8	24.2	22.1	37.3	85.8	**28.0**	**26.8**	40.7	57.2	38.3	37.8
RandSF.Q_{ssc}+SSync	**55.5**	**71.4**	**25.1**	**23.1**	**39.4**	**86.5**	27.9	**26.8**	**48.6**	**57.5**	**42.1**	**41.9**

are frequently decomposed into multiple slots, SSync achieves the best results by consolidating fragmented identities through selective alignment. SSync also yields substantial gains on MOVi-E, where the primary challenge is to accurately capture the boundaries of numerous small objects. Lastly, on YTVIS dataset, SSync achieves competitive FG-ARI while obtaining the highest mBO, indicating stronger object coverage. It is worth noting that YTVIS provides GT annotations only for a sparse set of salient foreground objects. Therefore, these metrics may not fully capture qualitative differences outside the evaluated objects. Accordingly, we provide qualitative comparisons in the Appendix; in shown examples, SSync appears to distinguish secondary objects and background regions more consistently than competing methods. Overall, these results suggest that selective alignment more effectively resolves encoder–decoder discrepancy than the dense alignment strategy, despite its simplicity, requiring no additional architectural components used by SRL [14].

Table 4: Image object-centric learning.

(a) Results on MOVi-E.

Method	FG-ARI ↑
VideoSAUR [10]	78.4
SOLV [45]	80.8
SlotContrast [21]	84.8
SlotCurri	**84.9**
SSync (Ours)	**86.0**

(b) Results on COCO.

Method	FG-ARI	mBO
Baseline	40.5	28.8
SRL [14]	42.8	**29.4**
SlotCurri [15]	**43.4**	28.9
SSync (Ours)	**47.9**	**33.1**

In addition, we integrate SSync into two variants of RandSF.Q [20] built upon VideoSAUR [10] and SlotContrast [21]. As reported in Tab. 3, SSync consistently improves performance, which indicates that SSync functions as a plug-and-play method, reinforcing its practical applicability.

Image Benchmarks. To evaluate generalization beyond video benchmarks, we assess SSync on image-level object-centric benchmarks. To illustrate, SSync achieves a SOTA FG-ARI of 86.0 on MOVi-E, and attains an ARI of 47.9 and mBO of 33.1 on real-world COCO2017, substantially outperforming SRL. This suggests that the proposed SSync improves spatial consistency independently of motion signals, further validating its robustness across diverse visual domains.

Table 5: Memory comparison (SRL/SSync). Values represent the maximum VRAM allocated per GPU in GB. OOM indicates Out-Of-Memory errors on an NVIDIA RTX PRO 6000 Blackwell GPU (97GB).

Frames (T)	Batch Size per GPU	
	32	64
$T = 4$	70 / **27**	OOM / **59**
$T = 6$	OOM / **48**	OOM / **89**
$T = 8$	OOM / **60**	OOM / **93**

Memory Efficiency. We compare the memory footprint against SRL [14] using mixed-precision (FP16) on YTVIS (518×518 resolution) in Tab. 5 (maximum VRAM per GPU). SRL ex-

hibits quadratic memory growth ($\mathcal{O}((T \cdot H \cdot W)^2)$) due to dense patch-to-patch comparisons, whereas SSync scales approximately linearly with the number of patches. As a result, at batch size 32 per GPU (T=4), SRL requires 70GB while SSync uses only 27GB, reducing memory usage by roughly 60% (and only a marginal 5% increase over SlotContrast [21]). Furthermore, SRL already encounters OOM at larger temporal lengths or batch sizes on our GPU, whereas SSync remains feasible. Note that per GPU VRAM values correspond to the maximum reserved memory, thereby not strictly linear with increasing T due to the properties of CUDA and cuDNN.

4.3 Ablation Study

All studies are conducted on MOVi-C.

Component Ablation. We analyze the contribution of each component in Tab. 6. Baseline model achieves 69.0 FG-ARI and 30.6 mBO. Applying boundary supervision ($\mathcal{L}_{\text{bd}}$) alone improves performance to 72.9 FG-ARI and 33.4 mBO. This confirms that selectively distilling sharp boundary cues from the encoder effectively refines decoder object maps. Similarly, interior supervision ($\mathcal{L}_{\text{nbd}}$) alone yields 71.4 FG-ARI and 33.3 mBO, indicating that decoder-guided denoising stabilizes noisy encoder assignments, leading to stable training. When both selective alignment losses are combined, performance increases substantially to 77.1 FG-ARI and 38.0 mBO, demonstrating that boundary calibration and interior denoising are complementary. Finally, incorporating transitive merging further improves performance to 79.4 FG-ARI and 39.5 mBO. This additional gain highlights the importance of stabilizing pseudo-label identities before cross-distillation. Overall, these results validate the selective alignment principle: region-aware supervision improves both boundary precision and slot consistency, while redundancy-aware merging ensures global semantic coherence.

Table 6: Component ablation study. $\mathcal{L}_{\text{bd}}$, $\mathcal{L}_{\text{nbd}}$, and T.M. represent calibrating the decoder boundary, denoising the attention map, and transitive pseudo-label merging, respectively.

Selected Components			MOVi-C	
$\mathcal{L}_{\text{bd}}$	$\mathcal{L}_{\text{nbd}}$	T.M.	FG-ARI	mBO
			69.0	30.6
✓			72.9	33.4
	✓		71.4	33.3
✓	✓		**77.1**	**38.0**
✓	✓	✓	**79.4**	**39.5**

Robustness to Number of Slots. A major challenge in video slot attention is over-fragmentation, which worsens when the number of slots (S) exceeds the actual objects in a scene. This forces prior works [14, 21] to carefully tune S per dataset to prevent multiple slots from undesirably cooperating to reconstruct a single object. As shown in Tab. 7, when prior approaches are trained with varying slot counts ($S \in \{7, 11, 15\}$), the performance deteriorates significantly at $S = 15$, indicating severe sensitivity to slot over-parameterization. In contrast, SSync maintains consistent performance across all configurations. This stability is driven by our transitive pseudo-label merging, which leverages spatio-temporal consistency to consolidate overlapping, fragmented slots into unified object representations. By inherently counteracting over-fragmentation, SSync

Table 7: Performance comparison with varying number of slots.

Method	slot=7		slot=11		slot=15	
	FG-ARI↑	mBO↑	FG-ARI↑	mBO↑	FG-ARI↑	mBO↑
SlotContrast [21]	74.9	27.9	69.3	32.7	61.8	**31.2**
SRL [14]	76.5	**31.6**	74.3	**34.5**	72.8	31.1
SlotCurri [15]	**78.7**	27.8	**77.6**	32.8	**74.8**	33.9
SSync (Ours)	**76.9**	**39.8**	**79.4**	**39.5**	**78.8**	**41.0**

drastically reduces sensitivity to the hyperparameter S, alleviating the burden of dataset-specific tuning and offering a highly practical framework for VOCL.

Transitive Pseudo-Label Merging. A key design choice in transitive merging is the source of active-region extraction used to identify redundant slots. While we derive active regions from the encoder attention map $\mathbf{A}$, we evaluate alternative formulations in Tab. 8 (a). Specifically, we compare: (a0) deriving active regions solely from the decoder object map $\mathbf{D}$; (a1–a2) hybrid logical criteria, where redundancy is independently computed from both $\mathbf{A}$ and $\mathbf{D}$ via Eq. 7–9 and combined using union (merge if either suggests redundancy) or intersection (merge only if both agree); and (a3) averaging IoU scores from $\mathbf{A}$ and $\mathbf{D}$ (Eq. 8 applied to each map) prior to thresholding with τ_{merge} in Eq. 9. Although using $\mathbf{D}$ as the merging source yields slightly lower performance, the overall results remain consistently strong across all configurations. This robustness suggests that redundancy detection reflects intrinsic spatio-temporal activation patterns, rather than dependence on a specific map representation.

Table 8: Variants of transitive merging strategies. A and D denote attention map and decoder object map, respectively.

	Strategy	FG-ARI↑	mBO↑
	Ours	79.4	39.5
	Varying criteria for merging		
(a0)	D	77.5	38.7
(a1)	A ∨ D	78.6	39.8
(a2)	A ∧ D	78.9	39.1
(a3)	Avg(A, D)	79.2	40.0
	Merging Range		
(b)	Pairwise	78.2	39.3
	Comparison to Slot Reg. [14]		
(c)	Slot Reg.	74.6	35.8

We also compare transitive merging with a pairwise baseline (Tab. 8(b)). Whereas ours constructs a connectivity graph over redundant slots and merges all slots within a connected component simultaneously, pairwise merging combines only the most similar slot pair whose IoU exceeds τ_{merge}. Results suggest that over-fragmented objects are often distributed across multiple slots, requiring global consolidation rather than pairwise agglomeration.

Finally, we compare our merging strategy with the slot regularization method used in SRL [14] (Tab. 8(c)). Slot regularization mitigates redundancy by penalizing overlapping slot pairs only during a predefined warm-up stage. In contrast, our method explicitly consolidates redundant slot identities based on spatio-temporal coverage similarity throughout the training. Empirical results demonstrate that transitive merging provides a more effective remedy for over-fragmentation, while eliminating the need for heuristic scheduling of regularization phases.

Impact of λ_{SSync}. λ_{SSync} controls the relative weight of the SSync loss against the base objective. As shown in Tab. 9a, SSync consistently improves performance over the baseline (69.0 FG-ARI, 30.6 mBO). Performance increases

Table 9: Hyperparameter analysis. The gray row denotes our default configuration.

(a) Impact of λ_{SSync}.

λ_{SSync}	FG-ARI	mBO
0.1	74.1	37.8
0.5	79.1	39.0
1.0	79.4	39.5
1.2	78.9	40.2

(b) Impact of n_{bd}.

n_{bd}	FG-ARI	mBO
1	79.4	39.5
2	78.9	39.3
3	78.2	39.4
4	76.6	36.7

(c) Impact of n_{nbd}.

n_{nbd}	FG-ARI	mBO
1	79.4	39.5
2	78.9	41.2
3	78.8	40.1
4	76.7	36.6

(d) Impact of τ_{merge}.

τ_{merge}	FG-ARI	mBO
0.65	78.3	40.1
0.7	79.4	39.5
0.75	79.1	39.5
0.8	78.2	39.0

steadily as λ_{SSync} grows until $\lambda_{\text{SSync}} = 1.0$, where selective alignment and reconstruction are well balanced. Beyond this point, further increasing the weight yields diminishing returns, indicating that excessive alignment may over-constrain the representation. Nevertheless, performance remains stable for slightly larger values (e.g., $\lambda_{\text{SSync}} = 1.2$), demonstrating robustness to moderate over-weighting.

Impact of n_{bd} and n_{nbd}. n_{bd} and n_{nbd} determine the sensitivity of boundary and interior region selection. Larger n_{bd} imposes stricter boundary criteria, while larger n_{nbd} relaxes interior selection. Since interior regions typically occupy the majority in most scenes, we fix $n_{\text{bd}} = n_{\text{nbd}} = 1$ to balance between the complementary supervision signals. Yet, our empirical results (Tab. 9b and 9c) confirm that performance remains stable across a reasonable range of values.

Impact of τ_{merge}. The merging threshold τ_{merge} controls the required spatio-temporal overlap for consolidating redundant slots. At the extremes, $\tau_{\text{merge}} = 1.0$ disables merging, while $\tau_{\text{merge}} = 0$ collapses all slots into a single identity. Tab. 9d shows that $\tau_{\text{merge}} = 0.7$ performs best on MOVi-C. Yet, the performance remains consistently high for thresholds around 2/3. This indicates that redundancy consolidation depends primarily on a clear overlap structure rather than precise threshold tuning, further validating the robustness of transitive merging.

4.4 Analysis

Analysis of the impact of denoising and deblurring. To quantify the impact in reducing isolated noise and boundary spillover, we report two diagnostics on MOVi-C in Tab. 10. First, the frame-averaged connected components (FCC) measures fragmentation: for each frame, we count the number of connected components in each per-slot binary mask and sum them over slots, then average across frames. A lower FCC indicates a reduction in isolated noise and over-fragmentation. SSync significantly suppresses spurious predictions, approaching the GT reference of 6.27. Second, we measure boundary

Table 10: Effects of encoder denoising and decoder deblurring on MOVi-C. FCC_8 measures mask fragmentation via the average number of connected components per frame, where connectivity is defined using an 8-neighborhood. GT FCC_8 is 6.27. Match_{90} reports the number of matched GT–slot pairs achieving at least 90% GT coverage. For these qualified pairs, we report the outside leakage (Leak) which indicates the total number of pixels assigned to the matched slot but lying outside the corresponding GT mask, summed over the spatio-temporal volume at the original video resolution ($\times 10^3$ pixels).

Method	Encoder	Decoder	
	$\text{FCC}_8 \downarrow$	$\text{Match}_{90} \uparrow$	Leak $\downarrow$
SlotContrast [21]	33.20	639	98.95
SRL [14]	21.03	684	86.26
SSync (Ours)	**8.79**	**702**	**72.02**

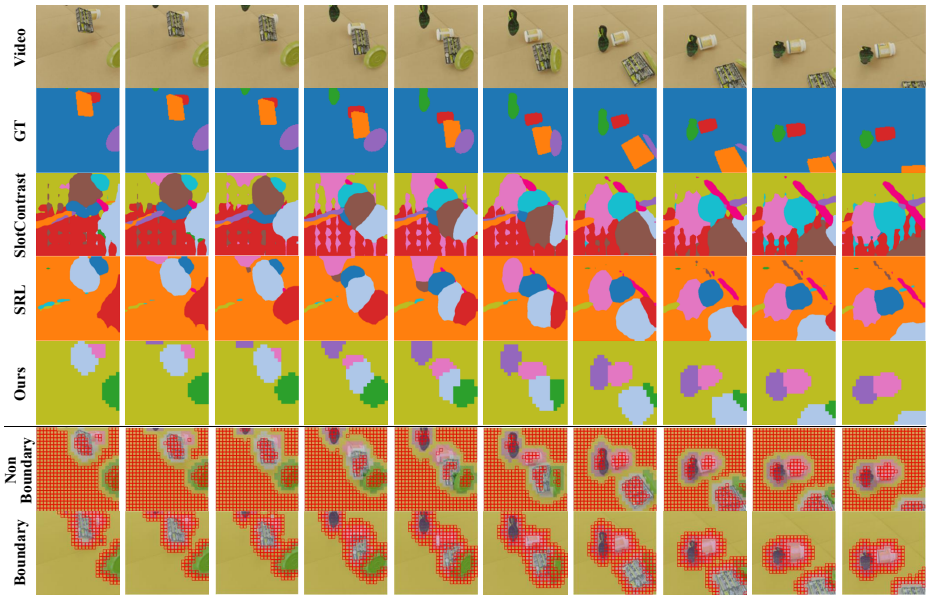

Fig. 3: Qualitative comparison on MOVi-C. From top to bottom, we visualize the input, GT masks, predictions from SlotContrast, SRL, and SSync, followed by the non-boundary and boundary regions selected by SSync for interior and boundary supervision, respectively.

spillover to assess spatial precision. We first match GT objects and predicted slots and retain only matched pairs whose GT coverage is at least 90% (MATCH_{90}). For these pairs, we quantify outside leakage as the number of predicted pixels lying outside the GT mask (lower indicates deblurred boundary). SSync not only yields more qualified matches than SRL, but also reduces the mean outside leakage per video by 16.5%. Collectively, these demonstrate that SSync excels at denoising fragmented assignments and sharpening slot coverage at object boundaries. Full results are in the Appendix.

Boundary Analysis. While mBO measures region overlap, it may not fully capture boundary sharpness. Thus, we evaluate boundary F-score [46] in Tab. 11. As shown, SSync improves F-score, supporting our claim that selective alignment leverages the encoder's boundary-sensitive cues to refine object contours.

Table 11: Boundary F-score on MOVi-C.

Method	F-score ↑
SlotContrast [21]	0.184
SRL [14]	0.222
SSync (Ours)	**0.255**

Qualitative Results. In Fig. 3, we present qualitative comparisons on MOVi-C. SlotContrast [21] exhibits severe noisy mask predictions, accompanied by spatially inflated and blurred object boundaries. Although SRL [14] mitigates part of this instability, residual noisy assignments and boundary ambiguity remain evident. In contrast, SSync produces more coherent object decomposition; boundaries are sharper and spatially consistent, while interior regions exhibit stable semantic grouping. Notably, background regions (although not evaluated by foreground-specific metrics like FG-ARI used in Tab. 2) are consistently

represented as unified and temporally coherent entities, while these regions are often fragmented across multiple slots in competing works. Also, we visualize the selected boundary and interior regions at the bottom of Fig. 3. Results illustrate that the model effectively separates spatio-temporal object transitions from interior regions across the video sequence. Additional qualitative comparisons and detailed region-selection visualizations are provided in the Appendix.

5 Conclusion

In this paper, we proposed Selective Synergistic Learning (SSync) for VOCL. SSync achieves impressive performance gains by both effectively and efficiently mitigating the discrepancy between the encoder's slot attention maps and the decoder's object maps. Specifically, SSync performs selective alignment between the attention map and object map only on regions where each map possesses its respective expertise, enabled by an efficient pseudo-labeling scheme. To further reduce the risk of pseudo-label corruption caused by object over-fragmentation, we introduced a transitive pseudo-label merging mechanism. By analyzing spatio-temporal slot activations and merging redundant slots via a connectivity-based criterion, we refine pseudo-label targets to be more semantically coherent and robust throughout training. Extensive experiments across VOCL benchmarks and additional evaluation protocols demonstrate the effectiveness of SSync, while remaining modular and easy to integrate into existing slot-based pipelines.

Acknowledgements

This work was supported in part by MSIT/IITP (No. RS-2022-II220680, RS-2020-II201821, RS-2019-II190421, RS-2024-00459618, RS-2024-00360227, RS-2024-00437633, RS-2024-00437102, RS-2025-25442569), MSIT/NRF (No. RS-2024-00357729), and KNPA/KIPoT (No. RS-2025-25393280).

References

1. Yanbo Wang, Letao Liu, and Justin Dauwels. Slot-vae: Object-centric scene generation with slot attention. In *International Conference on Machine Learning*, pages 36020–36035. PMLR, 2023.
2. 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.
3. Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. *arXiv preprint arXiv:2303.10834*, 2023.
4. Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Slot-vlm: Object-event slots for video-language modeling. *Advances in Neural Information Processing Systems*, 37:632–659, 2024.

5. Yerim Jeon, Miso Lee, WonJun Moon, and Jae-Pil Heo. Masking matters: Unlocking the spatial reasoning capabilities of llms for 3d scene-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 38668–38677, 2026.
6. Tatiana Zemskova and Dmitry Yudin. 3dgraphllm: Combining semantic graphs and large language models for 3d scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8885–8895, 2025.
7. Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. *Advances in Neural Information Processing Systems*, 37:113991–114017, 2024.
8. Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *The Twelfth International Conference on Learning Representations*, 2024.
9. 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.
10. Andrii Zadaianchuk, Maximilian Seitzer, and Georg Martius. Object-centric learning for real-world videos by predicting temporal feature similarities. *Advances in Neural Information Processing Systems*, 36:61514–61545, 2023.
11. 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.
12. Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional Object-Centric Learning from Video. In *International Conference on Learning Representations (ICLR)*, 2022.
13. 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.
14. Hyun Seok Seong, WonJun Moon, and Jae-Pil Heo. From vicious to virtuous cycles: Synergistic representation learning for unsupervised video object-centric learning. In *The Fourteenth International Conference on Learning Representations*, 2026.
15. WonJun Moon, Hyun Seok Seong, and Jae-Pil Heo. Reconstruction-guided slot curriculum: Addressing object over-fragmentation in video object-centric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.
16. Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification.
17. Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, and Francesco Locatello. Bridging the gap to real-world object-centric learning. In *The Eleventh International Conference on Learning Representations*, 2023.

18. Hongjia Liu, Rongzhen Zhao, Haohan Chen, and Joni Pajarinen. Metaslot: Break through the fixed number of slots in object-centric learning. *Advances in neural information processing systems*, 2025.
19. Rongzhen Zhao, Vivienne Huiling Wang, Juho Kannala, and Joni Pajarinen. Multi-scale fusion for object representation. In *The Thirteenth International Conference on Learning Representations*, 2025.
20. Rongzhen Zhao, Jian Li, Juho Kannala, and Joni Pajarinen. Predicting video slot attention queries from random slot-feature pairs. *arXiv preprint arXiv:2508.01345*, 2025.
21. 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*, pages 5401–5411, 2025.
22. Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.
23. Gilhan Park, WonJun Moon, SuBeen Lee, Tae-Young Kim, and Jae-Pil Heo. Mitigating background shift in class-incremental semantic segmentation. In *European Conference on Computer Vision*, pages 71–88. Springer, 2024.
24. Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Qian Zhang, and Wenyu Liu. Boxteacher: Exploring high-quality pseudo labels for weakly supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 3145–3154, 2023.
25. Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta, 2013.
26. Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.
27. Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in neural information processing systems*, 34:18408–18419, 2021.
28. Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.
29. Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11557–11568, 2021.
30. Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.
31. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
32. Robert M Haralick, Stanley R Sternberg, and Xinhua Zhuang. Image analysis using mathematical morphology. *IEEE transactions on pattern analysis and machine intelligence*, (4):532–550, 1987.
33. 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*, pages 23062–23071, 2024.
34. Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.
35. Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.
36. 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*, pages 3749–3761, 2022.
37. Tao Wang, Ning Xu, Kean Chen, and Weiyao Lin. End-to-end video instance segmentation via spatial-temporal graph neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10797–10806, 2021.
38. Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5188–5197, 2019.
39. Linjie Yang, Yuchen Fan, Yang Fu, and Ning Xu. The 3rd large-scale video object segmentation challenge-video instance segmentation track. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.(CVPR) Workshop*, 2021.
40. 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*, pages 740–755. Springer, 2014.
41. William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.
42. 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*, pages 2424–2433. PMLR, 2019.
43. Songtao Liu, Di Huang, and Yunhong Wang. Adaptive nms: Refining pedestrian detection in a crowd. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6459–6468, 2019.
44. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.
45. Görkay Aydemir, Weidi Xie, and Fatma Guney. Self-supervised object-centric learning for videos. *Advances in Neural Information Processing Systems*, 36:32879–32899, 2023.
46. Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016.