LEARNING WHEN TEMPORAL CONSISTENCY SHOULD BEGIN IN VIDEO OBJECT-CENTRIC LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised video object-centric learning must both discover new objects and preserve representations of objects already present in the scene. Slot–slot contrastive learning (SSC), a common temporal-consistency objective, favors the latter by encouraging each slot to match its previous-frame state. We show that stronger SSC improves representations of existing objects but progressively degrades representations formed for newly appearing objects, even when idle slot capacity remains available. This reveals an acquisition–maintenance trade-off in how temporal consistency is imposed. We introduce SLOTCOMMIT with learned-onset SSC, which learns a slot-specific onset for temporal consistency to support new-object acquisition without sacrificing the consistency of established representations. This onset is tied to a learned activity gate that determines whether a slot contributes to reconstruction. Upon first activation, the slot becomes committed, and its temporal-consistency constraint begins on the following transition. The first enabled SSC loss also trains the activation decision through commitment, coupling slot recruitment to a representation-dependent consistency cost. On matched MOVi-C models, learned-onset SSC increases acquisition of new objects by idle slots from 27.5% to 50.9% and improves simultaneous entrant–resident representation on separate slots from 25.5% to 33.2%. It also improves overall acquisition and both entrant and resident representation quality. Ablations show that both delaying the onset of SSC and training this onset through commitment contribute to the observed gains. On standard video grouping benchmarks, the complete SLOTCOMMIT model also improves over SlotContrast across MOVi-C, MOVi-E, and YouTube-VIS.

1 INTRODUCTION

Object-centric learning seeks structured visual representations that separate individual entities (Burgess et al., 2019; Greff et al., 2019). Such representations provide a basis for reasoning about object properties and interactions (Locatello et al., 2020). In video, this requires not only discovering objects in individual frames, but also maintaining object–representation correspondences as the scene evolves (Zhao et al., 2023; Wu et al., 2022). In the unsupervised setting, both object discovery and temporal correspondence must be learned from unlabeled videos, without annotations of object boundaries or identities across frames. A common approach is to propagate a set of latent representations, called *slots*, recurrently across frames and update them with new visual evidence (Kipf et al., 2021). Reconstruction encourages visual content to be grouped into slots, while recurrence carries their representations forward through time. When a new object enters the scene, the model must form a representation for it while preserving the object–slot correspondences already established for objects that remain visible. These two requirements place different demands on the slot representations. Existing objects benefit from continuity, whereas representing a new object requires some slot to change.

To promote continuity of existing representations, many video object-centric models use slot–slot contrastive learning (SSC) (Manasyan et al., 2025; Seong et al., 2026; Moon et al., 2026; Moon & Heo, 2026). SSC encourages each current slot to remain similar to its previous-frame state while distinguishing it from other slots. This pressure helps preserve established object–slot correspondences across frames but can interfere with forming new representations. A slot recruited for a newly appearing object must change from its previous state, while SSC encourages it to remain similar. This tension persists even when a low-utilization (*idle*) slot is available: although unused for the current

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

reconstruction, the slot still carries a recurrent state from earlier frames. The question is therefore not only how strongly temporal consistency should be enforced, but when it should begin for each slot.

We study this tension by distinguishing newly entering objects, which we call *entrants*, from objects already visible in the scene, which we call *residents*. On MOVi-C, adding SSC improves resident representations, while increasing its weight progressively lowers entrant quality, even when idle slots remain available (Fig. 1a). Reducing SSC globally eases acquisition but also weakens the temporal consistency that benefits established representations. Rather than choosing one consistency strength for both cases, we ask whether temporal consistency can begin at different times for different slots.

Motivated by this trade-off, we introduce SLOTCOMMIT with learned-onset SSC, a recurrent video slot-learning approach designed to accommodate newly entering objects while maintaining temporal consistency for established representations. Learned-onset SSC ties the onset of temporal consistency to each slot's learned activity. A binary gate determines whether the decoder uses a slot in the current frame. First activation establishes *commitment*, and the slot's own SSC term begins on the following transition, allowing it to first adapt to new content. Uncommitted slots still receive visual updates and can accumulate evidence before activation. The first SSC loss after commitment also backpropagates to the activation decision. This lets the gate account for the consistency cost of the representation it creates, while reconstruction rewards using a slot when it helps explain the frame. Together, these signals learn the onset of temporal consistency without annotations of object entry or slot activation. On the same architecture, learned-onset SSC raises idle-slot acquisition from 27.5% to 50.9% and joint entrant–resident representation from 25.5% to 33.2% on MOVi-C. It also improves total acquisition and both entrant and resident quality. Ablations show that delaying the first constrained transition and training the onset decision through commitment both contribute to these gains. The complete SLOTCOMMIT model further improves over SlotContrast on MOVi-C, MOVi-E, and YouTube-VIS.

Our contributions are **(1)** an acquisition-centered diagnosis showing an acquisition–maintenance trade-off in SSC around object entry; **(2)** learned-onset SSC, which ties each slot's consistency onset to its first learned activation and trains this onset through the first enabled SSC loss; and **(3)** controlled evidence that learned onset improves idle-slot recruitment and the simultaneous representation of newly entering and already established objects.

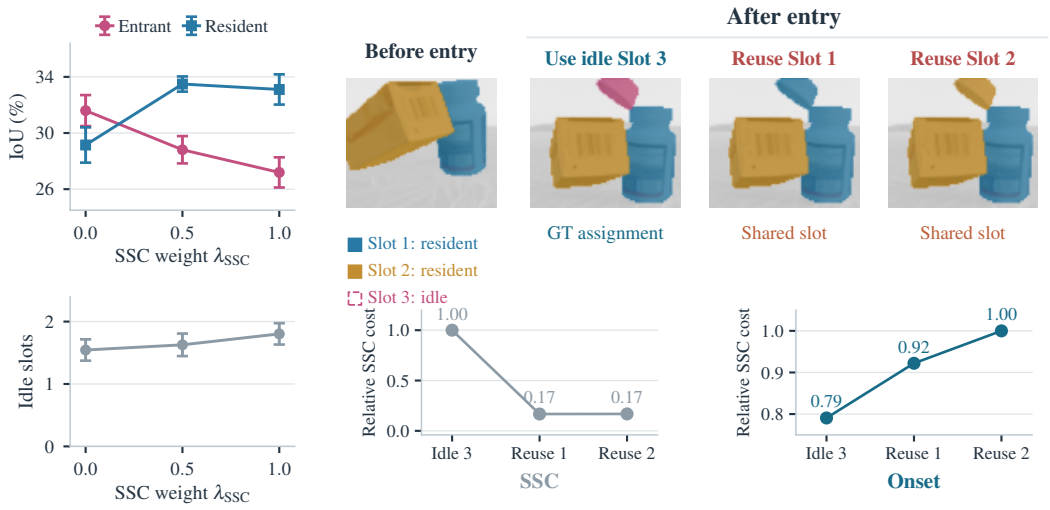
2 THE ACQUISITION–MAINTENANCE TRADE-OFF IN SSC

We begin with the recurrent slot setting of SlotContrast (Manasyan et al., 2025). At each frame, slots are updated from their previous states using current visual features and used to reconstruct the frame. SSC encourages each slot to remain similar to its previous state while distinguishing it from other slots, complementing reconstruction with temporal consistency. Object entry exposes the tension between maintaining and forming representations. Existing slots must preserve resident objects, while another slot must adapt to an entrant. Because SSC acts on these same recurrent transitions, entry events provide a direct setting for studying this trade-off. We therefore vary SSC strength in SlotContrast while keeping the remaining setup fixed.

2.1 MEASURING OBJECTS AT ENTRY

MOVi-C (Greff et al., 2022) provides the object masks and visibility information needed to localize entry events. We denote the first visible frame of a new object by t_e and call that object the *entrant*; objects already visible before t_e are *residents*. We evaluate 281 complete entry windows in which the entrant and at least one resident remain visible through $t_e + 3$. Annotations are used only for evaluation, with the full cohort construction given in Appendix B.

Our measurements capture formation and maintenance around the same event. *Entrant IoU* is the highest slot–object IoU at $t_e + 3$, measuring how well any slot represents the new object. For each resident, we select its best slot at $t_e - 1$ and evaluate the same slot at $t_e + 3$, then average over residents in each event. We call this *entry-local resident IoU*. Fixing the pre-entry slot prevents a different slot from receiving credit after the scene changes. We also measure whether low-utilization capacity exists before entry. A slot is *idle* when both its decoder area and mean attention are below 2% at $t_e - 1$. Such a slot contributes little immediately before entry but is not an empty state. It remains recurrent and carries a latent representation from preceding frames.



(a) SSC strength **(b)** Entrant assignments

Figure 1: **SSC strength and onset affect acquisition.** (a) Increasing SSC improves resident representation but reduces entrant quality, even though idle slot capacity remains available. (b) Three entrant assignments constructed from GT-pooled features. Under standard SSC, assigning the entrant to the idle Slot 3 incurs a higher consistency cost than reusing a resident slot in this example. Exempting Slot 3 from its own SSC term on the acquisition transition reverses this ordering while retaining the resident-slot terms. Costs are normalized within each rule. Together, the panels suggest that the difficulty is not a lack of slot capacity, but the temporal-consistency pressure applied before an idle slot has formed a new representation.

2.2 Resident Gains Accompany an Entrant Cost

We first ask whether stronger temporal consistency affects maintenance and formation differently around object entry. We train three seeds at each SSC weight, $\lambda_{\text{SSC}} \in \{0, 0.5, 1\}$, while holding the remaining setup fixed. The two groups respond differently as SSC becomes stronger. At weight 0.5, entry-local resident IoU improves from 29.1 to 33.5. Entrant IoU instead decreases as the SSC weight increases, falling from 31.6 without SSC to 27.2 at weight 1 (Fig. 1a). The full sweep in Appendix B.4 shows the same acquisition–maintenance trade-off.

The decline in entrant quality is not explained simply by an absence of available slots. Idle slots remain present throughout the SSC sweep. At the same time, the slot that best represents the entrant changes less from its previous state as SSC becomes stronger, as measured by cosine drift in Appendix B.4. Together, these observations point to a limitation beyond slot availability. An idle slot may be available in terms of utilization, but SSC still encourages its recurrent representation to remain close to its previous state. Acquiring an entrant therefore requires the slot not only to become utilized, but also to make a representation change that temporal self-matching can oppose. Figure 1b illustrates how this pressure can affect the preferred assignment. We construct three candidate assignments from object features while keeping the resident representations fixed. Under SSC, assigning the entrant to an idle slot incurs a higher consistency cost than sharing a resident slot in this example. When the idle slot is exempt from its own SSC term on the acquisition transition, the separated assignment instead receives the lower cost while the resident-slot terms remain active. Delaying the acquiring slot's SSC term can therefore reduce the consistency pressure on forming a new representation without removing temporal consistency from representations that are already established.

These results suggest that the central difficulty is not temporal consistency itself, but when that consistency begins to constrain a slot. Applying the same temporal pressure before a slot has formed a new representation can interfere with acquisition, while removing SSC globally would also weaken the pressure that benefits established representations. This motivates learning the onset of SSC separately for each slot, allowing a newly recruited slot to first adapt to new content before its own temporal-consistency term takes effect.

3 SLOTCOMMIT: LEARNING WHEN CONSISTENCY BEGINS

SLOTCOMMIT changes when SSC begins to constrain each slot. A learned activity gate determines whether a slot contributes to the current reconstruction. First activation establishes *commitment*, and the slot's own SSC term begins on the following transition rather than on the transition into activation. The first enabled SSC loss then backpropagates through commitment to the activation that established it, giving the onset decision a representation-dependent consistency cost alongside the reconstruction signal. We distinguish learned activity from the evaluation-time notion of idleness: a slot may receive recurrent updates while inactive, and an uncommitted slot is one that has never been active. The controlled SSC baseline uses the same recurrent architecture, evidence aggregation, and reconstruction gate, differing only in the SSC onset rule.

3.1 LEARNED-ONSET SSC

We build on the recurrent feature-reconstruction model of SlotContrast (Manasyan et al., 2025). Let Y_t denote frozen backbone features and $X_t = \phi(Y_t)$ their trainable projection. A predictor P propagates the previous slots, and a corrector $\mathcal{C}$ updates them using the current visual features,

$$S_t = \mathcal{C}(P(S_{t-1}), X_t), \qquad t > 1, \tag{1}$$

where $S_t = \{s_t^1, \ldots, s_t^K\}$ and the initial slots are learned.

SSC encourages these recurrent representations to remain temporally consistent. For a batch of B sequences with K slots each, we flatten the batch and slot indices into j, so j indexes all BK slots. Let $\bar{s}$ denote a normalized slot. The SSC loss matches each current slot to its own previous state while contrasting it against the previous states of the other slots,

$$\ell_{t,j}^{\mathrm{SSC}} = -\log \frac{\exp(\bar{s}_{t-1,j}^{\top} \bar{s}_{t,j}/\tau)}{\sum_{i=1}^{BK} \exp(\bar{s}_{t-1,i}^{\top} \bar{s}_{t,j}/\tau)}, \tag{2}$$

where τ is the temperature. Standard SSC applies this term to every slot on every transition.

Learned-onset SSC instead determines when each slot begins to incur its own SSC term. We associate slot k at frame t with binary activity z_t^k, indicating whether the decoder uses it in the current reconstruction. Activity history defines binary commitment,

$$c_t^k = \max(c_{t-1}^k, z_t^k), \qquad c_0^k = 0, \tag{3}$$

so c_t^k records whether the slot has activated by frame t. Its SSC term is enabled according to the previous commitment state,

$$\mathcal{L}_{\mathrm{SSC}}^{c} = \frac{1}{Q} \sum_{t=2}^{T} \sum_{j=1}^{BK} c_{t-1,j} \ell_{t,j}^{\mathrm{SSC}}, \qquad Q = B(T-1)K. \tag{4}$$

If a slot first activates at frame t_j^*, then $c_{t_j^*-1} = 0$, so its own SSC term is absent on the transition into activation. This sets $c_{t_j^*} = 1$, enabling SSC from the following transition. The slot can therefore first adapt to new content and is only then required to preserve the representation it has formed.

Learned onset changes only which slots incur their own self-matching terms. All slots remain in the contrastive candidate set in Eq. 2, including uncommitted slots, and the normalization Q remains fixed. Section 5.4 controls for the resulting change in effective SSC strength.

3.2 INFERRING ACTIVITY FROM CURRENT EVIDENCE

The activity gate uses visual evidence assigned to each slot during the final corrector iteration. For slot k, u_t^k summarizes the evidence content and m_t^k its total support. A shared gate maps (m_t^k, u_t^k) to binary activity z_t^k using a straight-through estimator during training. Appendix A.2 specifies the evidence aggregation and gate parameterization.

The decoder predicts a feature f_{tnk} and mask probability π_{tnk} for each token n and slot k, with activity controlling the slot's contribution,

$$\hat{y}_{tn} = \sum_{k=1}^{K} z_t^k \pi_{tnk} f_{tnk}. \tag{5}$$

When $z_t^k = 0$, the slot is removed from reconstruction without renormalizing the remaining mask probabilities, allowing reconstruction to evaluate whether using that slot improves the current frame.

Gating occurs after the recurrent update. An inactive slot therefore still receives current-frame visual evidence through the corrector and can accumulate support before activation. Activity indicates whether the slot is used for reconstruction, not whether it receives an update. It is also distinct from *idle* status in Sec. 2.1, which is an evaluation criterion based on observed utilization.

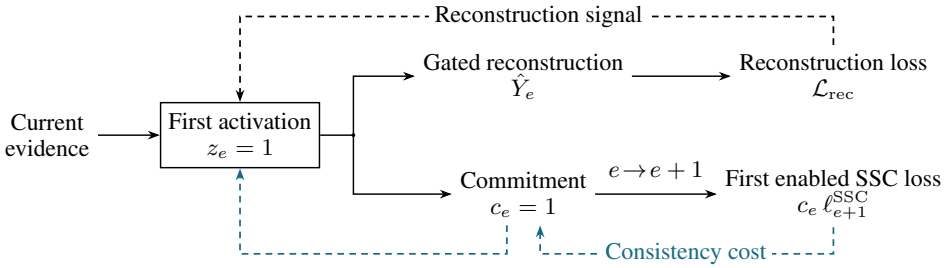


Figure 2: **Training signals for a learned SSC onset.** Activity controls both reconstruction and the onset of a slot's own SSC term. The transition into first activation is exempt from that term. From the following transition, SSC is enabled, and its first enabled loss backpropagates through commitment to the activation that established the onset. Reconstruction provides the complementary signal for using the slot when it helps explain the current frame.

3.3 Learning the Onset Decision

We jointly optimize reconstruction and learned-onset SSC,

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{SSC}} \mathcal{L}_{\text{SSC}}^{c}, \tag{6}$$

where $\mathcal{L}_{\text{rec}}$ is the mean-squared error between reconstructed features $\hat{Y}_t$ and frozen backbone features Y_t. Intuitively, reconstruction rewards activating a slot when it explains new evidence, whereas the first post-activation SSC term penalizes activations that lead to a temporally inconsistent representation.

Consider a slot whose first activation occurs at frame e. The transition into e is exempt because $c_{e-1} = 0$, while activation sets $c_e = 1$ and enables $c_e \ell_{e+1}^{\text{SSC}}$ on the following transition. The first enabled SSC loss can therefore backpropagate to the activation that established commitment, as illustrated in Fig. 2.

For binary forward values, commitment can equivalently be written as

$$c_t = 1 - \prod_{r \le t} (1 - z_r). \tag{7}$$

Our masked straight-through estimator differentiates through this expression while preventing disabled SSC terms from contributing a direct commitment gradient. At first activation,

$$\frac{\partial c_e}{\partial z_e} = \prod_{r < e} (1 - z_r) = 1. \tag{8}$$

Considering only the direct path through commitment and holding the slot representations fixed, the first enabled term $c_e \ell_{e+1}^{\text{SSC}}$ contributes ℓ_{e+1}^{SSC} to the gradient of the activation decision.

Because ℓ_{e+1}^{SSC} depends on temporal alignment relative to competing slots, this gives the activity decision a representation-dependent consistency cost. Reconstruction provides the complementary signal for activation when the slot helps explain the current evidence. Together, these signals learn when each slot begins to incur SSC without object-entry annotations or supervision over slot assignment. Section 5.4 tests this learning path by detaching commitment from activity.

4 Related Work

Object-centric learning from video. Recurrent slot models extend object-centric representations to video by propagating slots across frames. SAVi introduced recurrent slot propagation (Kipf

et al., 2021), while SAVi++ extended it to real-world videos with depth supervision (Elsayed et al., 2022). Transformer-based decoders were explored in SLATE and its video extension STEVE (Singh et al., 2021; 2022). More recent approaches reconstruct features from self-supervised encoders such as DINO (Caron et al., 2021), improving object discovery in images and video (Seitzer et al., 2022; Zadaianchuk et al., 2023). SlotContrast introduced slot–slot contrastive learning to encourage temporal consistency between slot representations (Manasyan et al., 2025). Our work instead studies when this temporal constraint should begin for each slot.

Grouping and slot allocation. Prior work has mainly addressed slot-allocation failures by changing how content is grouped across slots, rather than by changing when temporal regularization applies. SOLV merges similar slots (Aydemir et al., 2023), while SRL and SSync couple encoder and decoder grouping signals (Seong et al., 2026; Moon & Heo, 2026). SlotCurri adapts the slot set using reconstruction error and structural objectives (Moon et al., 2026). SRL, SSync, and SlotCurri retain SSC, whereas we focus on when SSC becomes active without introducing grouping objectives.

Activity and temporal state. SQAIR and SCALOR separate new-object discovery from propagation (Kosiorek et al., 2018; Jiang et al., 2019). DSSA separates appearance from identity across frames (Tran et al., 2026), while xSSC restructures temporal information through channel decomposition and cross-temporal reconstruction (Zhao et al., 2026). AdaSlot learns which slots participate in reconstruction (Fan et al., 2024). SLOTCOMMIT also learns slot activity, but uses first activation to determine SSC onset. The first enabled SSC loss then trains this decision through commitment, while uncommitted slots continue receiving recurrent visual updates.

5 EXPERIMENTS

We first examine how learning SSC onset changes entrant recruitment and representation quality, then test the onset rule and its learning signal through controlled interventions. Full-video grouping benchmarks evaluate the complete model across datasets.

5.1 SETUP

Controlled comparison. We compare learned-onset SSC with SSC on the same architecture, using identical evidence aggregation and decoder activity gates. Each slot's SSC term is weighted by c_{t-1} under learned-onset SSC and by 1 under SSC. Both train on MOVi-C with frozen DINOv2 ViT-S/14 features (Oquab et al., 2023), $K = 11$ slots, and $\lambda_{\mathrm{SSC}} = 0.5$. Appendix A.1 gives the shared recipe.

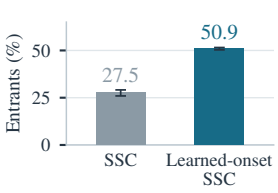
Acquisition and its source. We evaluate 322 entrants in 250 validation videos. Acquisition requires the same slot to have IoU > 0.3 for three consecutive visible frames, with the entrant as its best-matching foreground object. The first qualifying slot is the *acquisition source*, classified as idle or non-idle at $t_e - 1$ using Sec. 2.1. The *idle-slot acquisition rate* is the fraction of entrants ever acquired by an idle source, with unacquired entrants included in the denominator.

Quality and reporting. Entrant IoU uses all 281 complete windows, whether or not acquisition succeeds. We measure *full-video resident IoU* on 653 objects visible throughout their videos, selecting each object's slot with the highest mean IoU and keeping it fixed over time. This complements the entry-local resident measure. Tables report mean$\pm$sample SD over three matched training seeds unless stated otherwise.

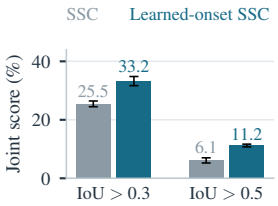
5.2 RECRUITING IDLE SLOTS FOR NEW OBJECTS

Learning SSC onset changes which slots acquire new objects. Idle-slot acquisition rises from 27.5% under SSC to 50.9% under learned-onset SSC (Fig. 3a), a 23.4-point gain compared with 5.5 points in total acquisition. The larger increase in idle-slot acquisition primarily reflects greater recruitment of idle slots rather than simply more successful acquisition.

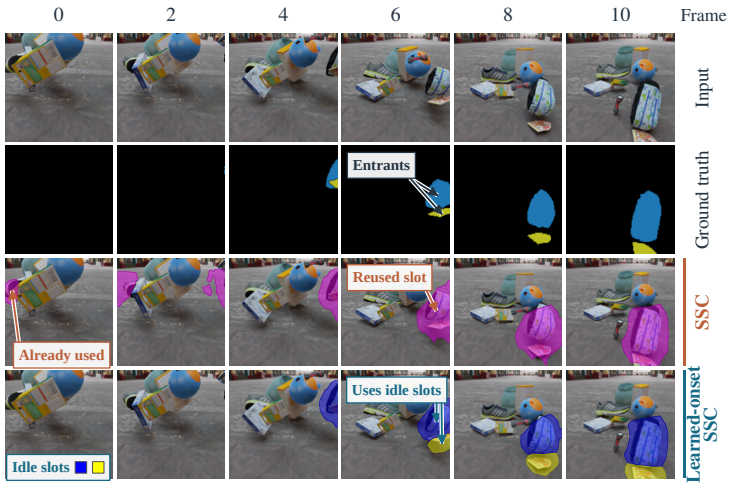
The increased recruitment of idle slots persists when we restrict attention to entrants acquired by both models. Among these shared successes, the fraction acquired from idle slots rises from 51.9% to 87.0%. The gain remains after removing differences in overall acquisition success. It is also robust to how acquisition and idleness are defined, holding across all twelve tested definitions in Appendix B.3.



(a) Idle-slot acquisition



(b) Joint representation



(c) Entrant acquisition

Figure 3: **Learned-onset SSC recruits idle slots for newly entering objects.** (a) Idle-slot acquisition as a percentage of all 322 entrants. (b) Joint entrant–resident representation over 281 complete windows. Bars show mean±SD over three seeds. (c) Two MOVi-C entrants. SSC reuses a previously active slot for one entrant; learned-onset SSC activates two idle slots and separates them. Only target objects are shown; colors identify slots within each model.

The qualitative examples in Fig. 3c show how slot recruitment differs. Under SSC, an entrant can be absorbed into a slot that was already representing another object. Learned-onset SSC instead activates unused slots and separates the new objects. The gate histories confirm that these acquisitions primarily come from newly activated slots: among idle-slot acquisitions under learned-onset SSC, 94.7% use slots that had not activated before entry. Appendix C gives the full analysis.

First activations occur around object entry. We next examine when slots are first brought into use, without conditioning on successful acquisition. Across all slots and 281 complete entry windows, the interval from t_e to $t_e + 3$ contains 1.85 ± 0.08 first activations per event under learned-onset SSC, compared with 1.12 under a permutation control matched by frame age and initial object-count group. Activations are concentrated at entry and the following frame, indicating that retained slot capacity is brought into use when new visual content appears. Because the analysis includes every slot, it also captures activations that do not lead to sustained acquisition. Appendix C.2 gives the full procedure and temporal profile.

5.3 Representing Entrants and Residents Together

Recruiting idle slots is useful only if it leads to better object representations rather than merely changing which slots become active. We therefore measure representation quality for both newly appearing and already established objects. Learned-onset SSC improves entrant IoU, full-video resident IoU, total acquisition, and FG-ARI on the matched architecture (Table 1).

Table 1: **Joint gains on a matched architecture.** MOVi-C, mean±SD over three seeds, 0–100 scale. Resident IoU is measured over full videos; entrant IoU includes every complete entry window.

Objective	Acquisition rate	Entrant IoU	Resident IoU	FG-ARI
SSC	52.7 ± 0.9	27.8 ± 0.8	40.6 ± 0.5	74.0 ± 0.4
Learned-onset SSC	$\mathbf{58.2} \pm 1.3$	$\mathbf{31.4} \pm 0.6$	$\mathbf{41.8} \pm 0.2$	$\mathbf{79.9} \pm 0.3$

Because entrant and resident quality can improve separately without both objects being represented at the same time, we also measure their simultaneous representation. For each resident, the *joint representation score* fixes its best slot before entry. At $t_e + 3$, an entrant–resident pair succeeds when

this fixed slot represents the resident above an IoU threshold and a different slot represents the entrant above the same threshold. We first average over resident pairs within each event and then across all 281 complete windows. Unlike the acquisition metric, this is an endpoint measure and does not require a three-frame hold.

The joint score increases from 25.5% under SSC to 33.2% under learned-onset SSC at IoU > 0.3 (Fig. 3b). This increase reflects better simultaneous representation while preserving the resident assignment established before entry. The gain holds at IoU > 0.5, with full results in Appendix C.3.

Following fixed slot assignments over time. The higher entrant IoU is not limited to the endpoint used by the joint score. If we select each entrant's best slot at the endpoint and follow that same slot throughout the entry window, its mean IoU increases from 19.3 to 22.1 under learned-onset SSC. Residents also show higher fixed-slot quality both before and after entry, with a similar average change across the entry event. Appendix C.4 gives these temporal diagnostics.

5.4 Learning the Commitment Decision

The preceding results show that learned-onset SSC changes recruitment and improves representation quality. We next isolate which parts of the rule are responsible. Table 2 tests three questions. We vary SSC strength to ask whether weaker temporal regularization reproduces learned onset, move the onset to the transition into first activity to test whether delaying the first constrained transition matters, and detach commitment to measure the contribution of training the activity decision through SSC.

Table 2: **Delayed onset and learning through commitment both matter.** MOVi-C on a shared architecture; mean±SD over paired seeds 42–44. The onset and gradient variants use $\lambda_{\mathrm{SSC}} = 0.5$.

SSC training rule	Idle-slot acquisition ↑	Entrant IoU ↑	Resident IoU ↑
SSC, $\lambda_{\mathrm{SSC}} = 0$	20.4 ± 3.4	$\mathbf{33.2} \pm 0.8$	37.5 ± 1.5
SSC, $\lambda_{\mathrm{SSC}} = 0.25$	28.1 ± 1.6	29.5 ± 0.6	40.2 ± 0.8
SSC, $\lambda_{\mathrm{SSC}} = 0.5$	27.5 ± 1.6	27.8 ± 0.8	40.6 ± 0.5
Include first activity, c_t	24.7 ± 2.1	21.7 ± 1.3	39.3 ± 0.9
Detached learned onset, c_{t-1}	28.9 ± 0.5	28.6 ± 0.3	41.2 ± 0.6
Learned-onset SSC, c_{t-1}	$\mathbf{50.9} \pm 0.6$	31.4 ± 0.6	$\mathbf{41.8} \pm 0.2$

Reducing SSC strength does not reproduce learned onset. The acquisition–maintenance trade-off remains on the matched gated architecture. Weakening SSC improves entrant quality but reduces resident quality, while none of the tested SSC weights approaches the idle-slot recruitment obtained by learned-onset SSC. Even removing SSC entirely raises entrant IoU above learned onset but lowers resident IoU to 37.5. A separate control that matches the average effective SSC weight of learned-onset SSC also yields substantially lower idle-slot acquisition (Appendix D.1). Learned onset therefore differs from simply reducing the overall amount of temporal regularization.

Delaying SSC beyond the first-activity transition is important. Replacing c_{t-1} with current commitment c_t enables SSC on the transition into first activity. Idle-slot acquisition then falls from 50.9% to 24.7%, while entrant IoU decreases from 31.4 to 21.7. In this control, constraining the first-activity transition substantially hurts acquisition.

Gradient flow from SSC to the onset decision is important. Detaching commitment preserves the same forward onset rule but removes the direct SSC gradient from commitment to activity. Idle-slot acquisition falls from 50.9% to 28.9%, while resident IoU changes only from 41.8 to 41.2. Thus, delaying SSC in the forward pass alone does not reproduce the full model's idle-slot recruitment. The strongest recruitment occurs when the first enabled SSC loss also trains the activity decision that determines its onset.

Increasing idle capacity alone does not reproduce the gain. Learned-onset SSC retains more idle slots before entry, so improved recruitment might simply reflect greater unused capacity. We test this explanation with two controls. First, we add an activity-sparsity penalty with warmup to standard SSC. This yields a pre-entry idle count close to learned-onset SSC, yet idle-slot acquisition reaches only 19.6% (Table 3).

Table 3: **Activity sparsity and idle-slot recruitment.** Mean±SD over three seeds. Idle capacity uses the same pre-entry utilization criterion in all rows.

Training objective	Idle slots	Idle-slot acquisition (%)
SSC	2.49 ± 0.32	27.5 ± 1.6
SSC + activity sparsity	5.18 ± 0.20	19.6 ± 1.9
Learned-onset SSC	5.69 ± 0.14	50.9 ± 0.6

Second, we intervene only at evaluation by supplying fresh queries whenever fewer than five existing slots are idle. This increases the observed pre-entry idle count of the SSC model to 5.54. Acquisition from slots that were already present and idle before entry reaches 29.3%, and rises only to 31.8% even when the newly inserted queries are counted as additional available sources. Increasing capacity either during training or at evaluation leaves a substantial gap to learned-onset SSC. In these controls, having more unused slots does not reproduce the 50.9% idle-slot acquisition rate; how those slots are brought into use also matters. Appendices D.4 and D.5 provide the implementation details.

5.5 Evaluation on Standard Video Grouping Benchmarks

The preceding experiments isolate learned-onset SSC on a matched architecture and focus on acquisition around object entry. We next evaluate the complete SlotCommit model on standard video grouping benchmarks across MOVi-C, MOVi-E, and YouTube-VIS. SlotCommit improves over SlotContrast in both FG-ARI and mBO across all three datasets (Table 4). We also report SSC-based methods with additional grouping or allocation mechanisms for context.

Table 4: **Video grouping across datasets.** Parenthetical annotations identify the temporal objective of methods in the first block; the second block reports SSC-based methods with additional grouping or allocation changes. SlotCommit uses 3 seeds.

	MOVi-C		MOVi-E		YouTube-VIS 2021	
Method	FG-ARI	mBO	FG-ARI	mBO	FG-ARI	mBO
Methods with different temporal objectives						
VideoSAUR (temporal-similarity prediction)	$64.8_{\pm 1.2}$	$38.9_{\pm 0.6}$	$73.9_{\pm 1.1}$	$35.6_{\pm 0.5}$	$28.9_{\pm 0.4}$	$26.3_{\pm 0.2}$
SlotContrast(SSC)	69.3	32.7	82.9	29.2	38.0	33.7
SlotCommit(learned-onset SSC)	$79.9_{\pm 0.3}$	$37.1_{\pm 0.4}$	$85.2_{\pm 0.1}$	$32.2_{\pm 0.4}$	$43.9_{\pm 0.5}$	$37.1_{\pm 0.1}$
SSC-based methods with grouping or allocation changes						
SRL	74.3	34.5	81.9	29.3	42.9	35.6
SlotCurri	$77.6_{\pm 0.9}$	$32.8_{\pm 0.2}$	$83.7_{\pm 0.2}$	$28.9_{\pm 0.7}$	$44.8_{\pm 1.2}$	$35.5_{\pm 2.2}$
SSync	$79.4_{\pm 0.6}$	$39.5_{\pm 0.1}$	$84.0_{\pm 0.9}$	$34.8_{\pm 1.9}$	$42.6_{\pm 0.2}$	$38.7_{\pm 0.6}$

6 Conclusion

We identified an acquisition–maintenance trade-off in SSC, in which temporal consistency helps preserve established object representations but can hinder the formation of new representations. SlotCommit addresses this trade-off with learned-onset SSC, allowing a slot to adapt before its own consistency term begins while using the first enabled SSC loss to train the onset decision through commitment. Matched experiments show improved idle-slot recruitment and simultaneous entrant–resident representation. The complete model also improves over SlotContrast across MOVi-C, MOVi-E, and YouTube-VIS.

Limitations. Our study focuses on diagnosing and modifying SSC in recurrent slot-based models with fixed slot budgets and frozen visual features. Extending learned onset to adaptive slot sets and combining it with complementary grouping or training techniques remains open.

AI USE STATEMENT

In this work, we used generative AI tools (i.e. ChatGPT) for revising passages, and correcting grammar. All AI-assisted text was reviewed by the authors. The research ideas, methodology, experimental execution, empirical results, and final scientific conclusions were developed and validated by the authors. The authors take full responsibility for the final submission.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A describe the model, losses, gradient estimators, architecture, and training procedures. Appendix B specifies the evaluation metrics, acquisition criterion, source-utilization measures, and sensitivity analyses. Appendix C provides additional recruitment and representation analyses, while Appendix D details the controlled interventions. We report variability across training seeds separately from video-bootstrap uncertainty for fixed checkpoints. Source code, training configurations, and trained models will be released upon publication.

REFERENCES

Görkay Aydemir, Weidi Xie, and Fatma Guney. Self-supervised object-centric learning for videos. *Advances in Neural Information Processing Systems*, 36:32879–32899, 2023.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

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

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.

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.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23062–23071. IEEE, 2024.

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, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3749–3761, 2022.

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

Jindong Jiang, Sepehr Janghorbani, Gerard De Melo, and Sungjin Ahn. Scalor: Generative world models with scalable object representations. *arXiv preprint arXiv:1910.02384*, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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.

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.

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.

Anna Manasyan, Maximilian Seitzer, Filip Radovic, Georg Martius, and Andrii Zadaianchuk. Temporally consistent object-centric learning by contrasting slots. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5401–5411. IEEE, 2025.

WonJun Moon and Jae-Pil Heo. Selective synergistic learning for video object-centric learning. In *European Conference on Computer Vision*, pp. 21–39. Springer, 2026.

WonJun Moon, Hyun Seok Seong, and Jae-Pil Heo. Reconstruction-guided slot curriculum: Addressing object over-fragmentation in video object-centric learning. *arXiv preprint arXiv:2603.22758*, 2026.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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.

Hyun Seok Seong, WonJun Moon, and Jae-Pil Heo. From vicious to virtuous cycles: Synergistic representation learning for unsupervised video object-centric learning. *arXiv preprint arXiv:2602.03390*, 2026.

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

Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. *Advances in neural information processing systems*, 35:18181–18196, 2022.

Sieu Tran, Duc Nguyen, Hao Vo, Khoa Vo, and Ngan Le. Dual-state slot attention: Decoupling appearance and identity for video object-centric learning. *arXiv preprint arXiv:2606.12601*, 2026.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models. *arXiv preprint arXiv:2210.05861*, 2022.

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.

Rongzhen Zhao, Zhiyuan Li, Juho Kannala, and Joni Pajarinen. Internalizing temporal consistency in video object-centric learning without explicit regularization. *arXiv preprint arXiv:2605.31508*, 2026.

Zixu Zhao, Jiaze Wang, Max Horn, Yizhuo Ding, Tong He, Zechen Bai, Dominik Zietlow, Carl-Johann Simon-Gabriel, Bing Shuai, Zhuowen Tu, et al. Object-centric multiple object tracking. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16555–16565. IEEE, 2023.

APPENDIX

A Implementation Details **14**

A.1 Architecture and Recipes 14

A.2 Evidence Aggregation and the Decoder Gate 14

A.3 Gradient Estimator 15

A.4 Seeds and Checkpoints 15

B Evaluation and Diagnosis **15**

B.1 Object Quality and Acquisition 16

B.2 Source Utilization and Reporting 16

B.3 Sensitivity to Acquisition and Idle Definitions 16

B.4 SSC Strength Sweep 16

C Acquisition and Representation **17**

C.1 Recruitment and Activity History 17

C.2 Activation Timing around Entry 17

C.3 Joint Representation Score 18

C.4 Quality over Time 18

D Control Details **18**

D.1 SSC Strength 18

D.2 Onset at First Activity 19

D.3 Gradient Detachment 19

D.4 Activity Sparsity 19

D.5 Query Replenishment 19

E Additional Qualitative Examples **20**

A Implementation Details

This appendix specifies the architecture, sampling rules, gradient estimators, and training recipes. Scores and rates use the 0–100 scale; probabilities, coefficients, slot counts, and cosine drift keep their native units.

A.1 Architecture and Recipes

SlotCommit builds on the recurrent SlotContrast pipeline with a decoder activity gate and learned-onset SSC. The SSC term in Eq. 2 has the InfoNCE form (Oord et al., 2018), using the same slot across consecutive frames as the positive pair. The shared backbone consists of frozen DINOv2 features (Oquab et al., 2023), a trainable feature projection, learned initial slots, a one-block four-head transformer predictor, and an iterative slot corrector that runs three iterations on the first frame and two afterwards. The decoder is a three-hidden-layer MLP of width 1024 that outputs a feature prediction and a mask logit per slot and token. Every slot receives every corrector update. Table 5 lists the per-dataset recipes.

Table 5: **Recipes.** All use batch size 64, SSC temperature 0.1, and $\lambda_{\mathrm{SSC}} = 0.5$. Resolution is square; steps are the scheduled training budget.

Dataset	DINOv2	Slot dim.	K	Resolution	Steps	Peak LR
MOVi-C	ViT-S/14	64	11	336	100k	4×10^{-4}
MOVi-E	ViT-B/14	128	15	336	300k	4×10^{-4}
YouTube-VIS 2021	ViT-B/14	64	7	518	100k	8×10^{-4}

The matched SSC/learned-onset study trains with Adam (Kingma & Ba, 2014) for 100k steps, using 2,500 warm-up steps, cosine decay to 4×10^{-7}, bfloat16 autocast, and gradient-norm clipping at 0.05. The ViT-S/14 backbone yields 384-dimensional features on a 24×24 grid. Training preprocessing uses bicubic RGB resizing, random horizontal flips, and ImageNet normalization; validation runs on full videos with nearest-exact resizing of the segmentation labels.

The gate head has hidden width 64 and is initialized with zero weights and bias 2.0. Gate and ownership relaxation temperatures are 1.0. At evaluation, the model runs over the full video with its recurrent corrector and gated decoder; commitment is not required because it is used only to determine the onset of the training-time SSC loss.

A.2 Evidence Aggregation and the Decoder Gate

DSSA identifies a token-renormalization artifact in which weak competitive attention is amplified during aggregation (Tran et al., 2026). We use exclusive token assignment as an alternative way to mitigate this effect, masking renormalized attention before aggregation. Both controlled models use this implementation.

For token n and slot k, attention competes over slots and is then normalized over tokens,

$$a_{tnk} = \mathrm{softmax}_k\Big((q_t^k)^\top k_t^n/\sqrt{d}\Big), \qquad r_{tnk} = \frac{a_{tnk}}{\sum_{n'} a_{tn'k} + \epsilon}, \tag{9}$$

with key dimension d and a small ϵ. A hard Gumbel–softmax assignment (Jang et al., 2016) h_{tnk} masks these weights,

$$h_{tnk} = \mathbb{I}\Big[k = \arg\max_j(\log a_{tnj} + g_{tnj})\Big], \qquad w_{tnk} = r_{tnk}h_{tnk}, \tag{10}$$

$$u_t^k = \sum_n w_{tnk} v_t^n, \qquad m_t^k = \sum_n w_{tnk}, \tag{11}$$

where g_{tnj} are Gumbel samples and v_t^n is the token value. Assignment uses a straight-through backward pass. The evidence vector u_t^k drives slot refinement, while m_t^k measures support; both are zero when no tokens are assigned. The gate head G_θ receives (m_t^k, u_t^k) from the final corrector iteration.

The gate uses a binary Concrete relaxation (Maddison et al., 2016). With logistic noise η, temperature τ_z, sigmoid σ, and stop-gradient sg,

$$p_t^k = \sigma\big((G_\theta(m_t^k, u_t^k) + \eta_t^k)/\tau_z\big), \qquad \hat{z}_t^k = \mathbb{I}[p_t^k > 0.5], \qquad z_t^k = \hat{z}_t^k + (p_t^k - \mathrm{sg}(p_t^k)). \tag{12}$$

The forward activity value is binary, while the backward pass uses a straight-through estimator (Bengio et al., 2013). During training, the gate samples noise; at evaluation, gate noise is disabled.

The gate multiplies the decoded mixture after the mask softmax and never changes the recurrent state update. Inactive contributions are removed without renormalizing the remaining mask probabilities, allowing reconstruction to evaluate the effect of removing a slot without automatically rescaling the surviving contributions. The matched SSC baseline uses the same evidence aggregation and decoder.

A.3 Gradient Estimator

Section 3.3 shows how the first enabled SSC loss backpropagates through commitment to first activation. Here we give the masked estimator and derive this direct gradient.

For binary forward gates, commitment is evaluated as in Eq. 3 and differentiated through the equivalent product form in Eq. 7. Each SSC summand is implemented as

$$\mathrm{where}(c_{t-1,j} > 0.5, \ell_{t,j}^{\mathrm{SSC}}, 0)\, c_{t-1,j}. \tag{13}$$

The Boolean comparison is not differentiated. Writing its value as $b_{t,j} = \mathbf{1}[c_{t-1,j} > 0.5]$, one enabled term contributes

$$\frac{\lambda_{\mathrm{SSC}}}{Q}\, b_{t,j} \ell_{t,j}^{\mathrm{SSC}} \frac{\partial c_{t-1,j}}{\partial z_{r,j}}, \qquad \frac{\partial c_{t-1,j}}{\partial z_{r,j}} = \prod_{\substack{s<t \\ s\neq r}} (1 - z_{s,j}), \quad r < t. \tag{14}$$

The total direct gradient at gate $z_{r,j}$ is

$$g_{r,j}^{\mathrm{SSC}} = \frac{\lambda_{\mathrm{SSC}}}{Q} \sum_{t=r+1}^{T} \mathbf{1}[c_{t-1,j} > 0.5]\, \ell_{t,j}^{\mathrm{SSC}} \prod_{\substack{s<t \\ s\neq r}} (1 - z_{s,j}). \tag{15}$$

Let $r < T$ be the first active frame. On the following transition, $c_{r,j} = 1$ and $\partial c_{r,j}/\partial z_{r,j} = 1$, so the first enabled SSC loss contributes $(\lambda_{\mathrm{SSC}}/Q)\ell_{r+1,j}^{\mathrm{SSC}}$ to the gradient at that activation. This contribution remains even if $z_{r+1,j} = 1$, since the term depends on history only through frame r. For any later summand in Eq. 15, the direct path is nonzero only while r is the sole active frame in its history. Once that history includes a second active frame, its product is zero; earlier terms still contribute to the total gradient. The mask gives no direct gate gradient for an all-zero history.

Let $\alpha_{r,j}$ denote the gate logit. The sigmoid straight-through estimator gives

$$\frac{\partial z_{r,j}}{\partial \alpha_{r,j}} = \frac{p_{r,j}(1 - p_{r,j})}{\tau_z} \geq 0, \tag{16}$$

so multiplying $g_{r,j}^{\mathrm{SSC}}$ by this derivative preserves the nonnegative sign of the direct gradient and its downward pressure on first activation. The derivation holds slot representations fixed and describes the direct commitment path; shared model parameters also receive reconstruction and representation gradients. Commitment is accumulated in float32. Reconstruction MSE averages over batch, frames, tokens, and channels.

A.4 Seeds and Checkpoints

All three-seed experiments use training seeds 42, 43, and 44, matched across compared conditions. We report the final scheduled checkpoint for every run, without validation-based checkpoint selection. The average-weight-matched control in Appendix D.1 uses one checkpoint.

B Evaluation and Diagnosis

All annotations are used only for evaluation. With zero-based frames, an entrant first reaches 32 annotated pixels at $t_e \geq 2$. Complete windows require the entrant and at least one resident to remain visible through $t_e + 3$. There are 281 such windows, while acquisition rates use all 322 entrants (Table 6).

Table 6: **MOVi-C evaluation populations.** Counts are objects or events on the 250 validation videos.

Population	Count	Use
All entrants	322	Acquisition rate and acquisition source
Complete windows with residents	281	Matched entrant IoU and the diagnosis; 170 videos
Full-video residents	653	Fixed-slot temporal-mean IoU

B.1 Object Quality and Acquisition

Entrant IoU is the maximum slot IoU at $t_e + 3$. Its first-four variant averages the per-frame maximum over $t_e, \ldots, t_e + 3$. Entry-local resident IoU fixes each resident's best slot at $t_e - 1$ and reads it at $t_e + 3$, averaging residents within events. Full-video resident IoU instead chooses the slot with highest mean IoU over the object's complete trajectory. The joint representation score is defined in Appendix C.3.

Acquisition requires three consecutive visible frames with the same slot above IoU 0.3 and the entrant as its best-matching foreground object. Its onset is the first frame of the first completed hold. Ties go to higher onset-frame IoU, then smaller slot index. A hold cut short by clip end does not qualify.

B.2 Source Utilization and Reporting

The first acquiring slot is the source. Its idle label is fixed at $t_e - 1$, when decoder area and mean attention must both be below 2%. Idle-slot acquisition divides by all entrants; the common-acquisition share restricts both models to entrants acquired by both. Slot indices need not correspond across models.

For the matched models, we bilinearly upsample the gated decoder masks $z_t^k \pi_{tnk}$ to the evaluation grid of $H \times W$ pixels (`align_corners=False`), then take the argmax over slots at each pixel to obtain labels $\hat{y}_t(p)$, breaking ties by the smallest slot index. Decoder area D_t^k is the resulting hard-mask area fraction; mean attention A_t^k averages the final corrector iteration's slot-softmax weights a_{tnk} from Eq. 9 over the N feature tokens, before token normalization and assignment masking:

$$D_t^k = \frac{1}{HW} \sum_{p=1}^{HW} \mathbf{1}[\hat{y}_t(p) = k], \qquad A_t^k = \frac{1}{N} \sum_{n=1}^{N} a_{tnk}. \tag{17}$$

Thus decoder area measures a fraction of pixels, while mean attention measures the average share of token attention assigned to the slot; the idle criterion is $D_{t_e-1}^k < 0.02$ and $A_{t_e-1}^k < 0.02$.

Metrics are computed per checkpoint, then averaged equally over three seeds (mean±sample SD). Paired differences compare models trained with the same seed. Video-bootstrap intervals use 3,000 whole-video resamples, keeping model pairs and video weights across seeds together, with checkpoints fixed. They quantify evaluation-set uncertainty; seed SDs quantify training variation. FG-ARI and mBO use full videos. Scores are multiplied by 100; thresholds remain in $[0, 1]$.

B.3 Sensitivity to Acquisition and Idle Definitions

The idle-slot acquisition gain is positive across all twelve tested combinations of utilization rule, IoU threshold, and hold duration (Table 7).

B.4 SSC Strength Sweep

The SlotContrast weight sweep uses the same 281 windows and three seeds per weight. Table 8 reports the complete means and SDs. Cosine drift is $1 - \cos(s_{t-1}^k, s_t^k)$, averaged over the four transitions from $t_e - 1$ to $t_e + 3$. We use the entrant's best slot at $t_e + 3$ and each resident's fixed pre-entry slot.

Table 7: **Idle-slot acquisition gain under alternative definitions.** Learned-onset SSC minus SSC in percentage points, paired over three seeds. IoU thresholds in native units.

Utilization rule	Acquisition IoU	Hold frames	Idle-slot acquisition gain
Joint 2%	0.3	3	23.4 ± 2.1
Joint 2%	0.3	1	27.1 ± 4.6
Joint 2%	0.5	3	12.4 ± 0.5
Joint 2%	0.5	1	20.8 ± 0.6
Decoder 1%	0.3	3	21.0 ± 1.6
Attention 1%	0.3	3	31.6 ± 3.4
Joint 1%	0.3	3	31.7 ± 3.2
Decoder 2%	0.3	3	16.9 ± 1.7
Attention 2%	0.3	3	22.9 ± 2.2
Decoder 5%	0.3	3	8.9 ± 2.6
Attention 5%	0.3	3	10.1 ± 1.6
Joint 5%	0.3	3	10.6 ± 1.6

Table 8: **SlotContrast response to SSC weight.** Mean±sample SD over three training seeds on the same 281 entry windows. IoUs in percent; slot counts and cosine drift unscaled.

Metric	$\lambda_{\mathrm{SSC}} = 0$	$\lambda_{\mathrm{SSC}} = 0.5$	$\lambda_{\mathrm{SSC}} = 1$
Entrant endpoint IoU	31.6 ± 1.1	28.8 ± 1.0	27.2 ± 1.1
Resident endpoint IoU	29.1 ± 1.3	33.5 ± 0.5	33.1 ± 1.1
Pre-entry idle slots	1.54 ± 0.17	1.63 ± 0.18	1.80 ± 0.17
Entrant cosine drift	0.0624 ± 0.0077	0.0377 ± 0.0009	0.0270 ± 0.0009
Resident cosine drift	0.0292 ± 0.0041	0.0259 ± 0.0008	0.0183 ± 0.0007

C Acquisition and Representation

C.1 Recruitment and Activity History

For the 322 entrants in 177 videos, we reconstruct each slot's observed activity history from the start of the evaluation clip, $c_{t,k}^{\mathrm{prefix}} = \bigvee_{\tau=0}^{t} z_{\tau,k}$. This prefix is an evaluation diagnostic; inference uses the recurrent model and activity gate. Of idle-slot acquisitions under learned-onset SSC, $94.71 \pm 0.77\%$ use slots with no activation before entry.

Among entrants acquired by both models, the idle-source share rises from $51.85 \pm 2.14\%$ under SSC to $87.00 \pm 2.17\%$ under learned-onset SSC. The higher idle-source share therefore persists when comparing the same acquired objects.

C.2 Activation Timing around Entry

We use the 281 GT-defined complete entry windows and count first activations across all slots, without conditioning on successful acquisition. A first activation occurs when $z_{t,k} = 1$ and $c_{t-1,k}^{\mathrm{prefix}} = 0$. Activity at initialization is included in the prefix history but is not counted as an activation transition. For each event, we sum first activations from t_e through $t_e + 3$ and then average across events.

Learned-onset SSC produces 1.85 ± 0.08 first activations per event from t_e to $t_e + 3$, compared with a permutation mean of 1.12. As shown in Fig. 4, activations concentrate at entry and the following frame rather than being distributed uniformly around the event.

The permutation control exchanges complete activation profiles across validation videos within groups defined by the initial GT object count, using groups < 4, 4–6, and ≥ 7. Each event is evaluated at its original frame indices in the donor profile, preserving frame age. All 250 validation videos form the donor pool. The permutation baseline is estimated from 10,000 random permutations using shared mappings across the three training seeds.

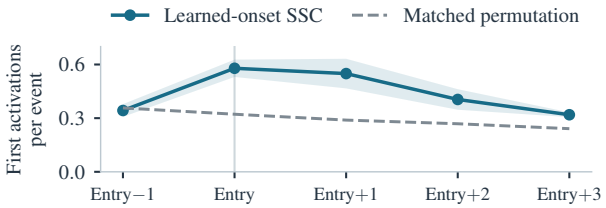


Figure 4: **First activations concentrate around object entry.** Per-frame first-activation counts across the 281 complete entry windows, including $t_e - 1$. Shading shows seed SD and the dashed line shows the matched permutation mean.

C.3 Joint Representation Score

For event e, let o_e be the entrant and R_e the continuously visible residents. We fix each resident r to its best slot before entry, $a_e(r)$. With $I_t(k, o)$ denoting IoU between slot k and object o, the joint score at threshold q is

$$J_e(q) = \frac{1}{|R_e|} \sum_{r \in R_e} \mathbf{1}[I_{t_e+3}(a_e(r), r) > q] \mathbf{1}\left[\max_{k \neq a_e(r)} I_{t_e+3}(k, o_e) > q\right]. \tag{18}$$

We average residents within each event, then all 281 complete events. Resident eligibility depends only on visibility, with no minimum pre-entry IoU. Each pair is scored independently at the endpoint; acquisition is defined separately by a three-frame hold.

Table 9: **Joint representation at $t_e + 3$.** Mean±sample SD over three seeds; differences are percentage points. Intervals use 3,000 paired whole-video bootstrap draws with fixed checkpoints.

Threshold	SSC	Learned-onset SSC	Difference [95% CI]
IoU > 0.3	25.46 ± 0.98	33.24 ± 1.57	$+7.78$ [5.46, 10.28]
IoU > 0.5	6.12 ± 0.88	11.22 ± 0.46	$+5.10$ [3.27, 7.07]

C.4 Quality over Time

Fixing the entrant's endpoint-best slot over all four frames gives IoU 19.3 ± 0.7 under SSC and 22.1 ± 0.4 under learned-onset SSC. Allowing the best slot to vary by frame gives 21.2 ± 0.5 and 24.4 ± 0.5. The gain therefore persists when the same slot is followed across the window.

Resident fixed-slot IoU is 45.03 ± 0.69 versus 47.86 ± 0.73 before entry and 36.61 ± 0.67 versus 39.60 ± 0.71 afterwards. The pre-to-post changes are -8.42 ± 1.34 and -8.26 ± 0.41, with a paired difference of 0.16 ± 1.76. Residents are represented better at both endpoints, with similar changes across entry. These measurements use all complete windows rather than conditioning on acquisition success.

D Control Details

D.1 SSC Strength

The sweep in Table 2 keeps the gated architecture and training schedule fixed, with three seeds per weight. Under learned-onset SSC, the enabled-term fraction is $\rho = Q^{-1} \sum_{t,j} c_{t-1,j}$, giving an average weight per possible term of $\lambda_{\mathrm{SSC}}\rho$. Over the last 10k steps, $\rho = 0.54$. A single-checkpoint SSC control at the corresponding weight 0.27 reaches 26.09% idle-slot acquisition.

D.2 Onset at First Activity

The onset control replaces c_{t-1} with c_t in Eq. 4, including the transition into first activity. This substitution changes the commitment history used by both the forward loss and the product estimator. Table 2 compares the two formulations across paired seeds 42–44.

D.3 Gradient Detachment

The detached control replaces $c_{t-1,j}$ with $\mathrm{sg}(c_{t-1,j})$ in Eq. 4. It preserves the forward gating rule and removes the direct SSC gradient through commitment, while reconstruction continues to train activity. Table 2 reports the three-seed comparison.

D.4 Activity Sparsity

Starting from the trained SSC baseline for each seed, we fine-tune for a further 100k steps with an AdaSlot-style penalty on decoder activity z (Fan et al., 2024). We retain reconstruction and SSC and introduce the activity penalty with warmup, using $\lambda_{\mathrm{sparse}} = 0.1$. Table 3 compares pre-entry idle capacity and idle-slot acquisition using the definitions in Appendix B.2. The penalty produces more idle slots while reducing their recruitment for entrants.

D.5 Query Replenishment

We also test adding queries to trained checkpoints at evaluation. After each frame, if fewer than five slots satisfy the idle criterion, we append enough copies of learned initial queries to bring the combined count of idle and fresh queries to five before the next frame. Copies cycle through the learned initial slots, while existing slots retain their states. The recurrent model then updates all queries with the next frame's evidence.

Replenishment increases the SSC model's mean observed pre-entry idle count from 2.49 ± 0.3 to 5.54 ± 0.1. Acquisition from sources already present and idle before entry reaches $29.3 \pm 1.7\%$. Counting queries inserted immediately before entry as additional available sources gives $31.8 \pm 1.6\%$. Table 10 compares native checkpoints and replenishment under this evaluation protocol.

Table 10: **Query replenishment at evaluation.** Mean$\pm$sample SD over three seeds. Idle-slot acquisition counts sources present and idle at $t_e - 1$; the text additionally accounts for queries inserted immediately before entry.

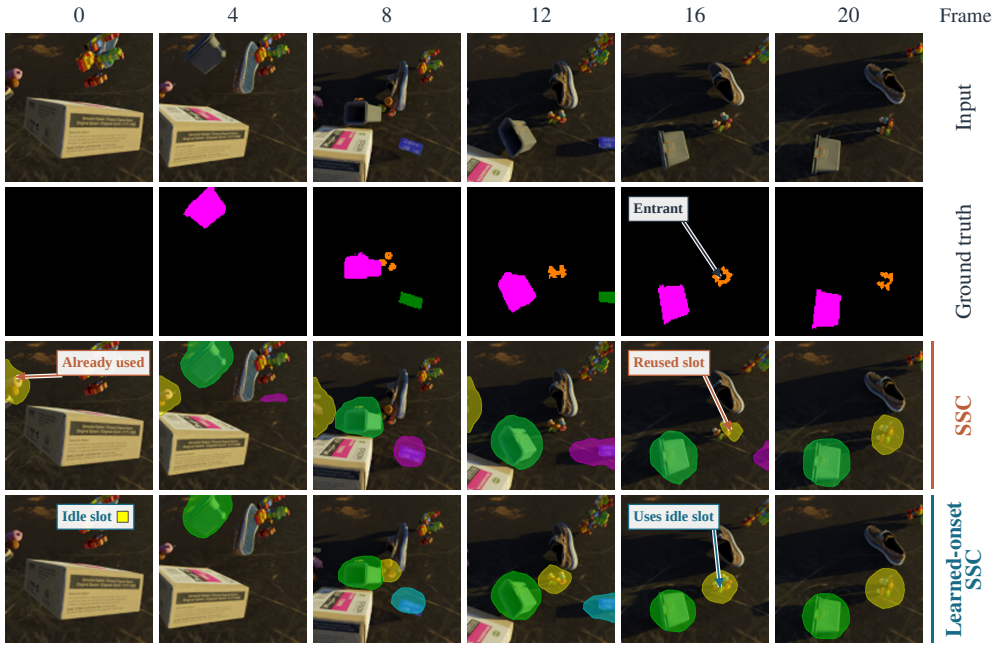
Variant	Idle-slot acquisition (%)	Entrant IoU	Resident IoU
SSC	27.5 ± 1.6	27.8 ± 0.8	40.6 ± 0.5
SSC + replenishment	29.3 ± 1.7	29.6 ± 0.8	37.1 ± 0.7
Learned-onset SSC	50.9 ± 0.6	31.4 ± 0.6	41.8 ± 0.2

E ADDITIONAL QUALITATIVE EXAMPLES

Figures 5 and 6 show additional examples on MOVi-C/E and YouTube-VIS 2021.

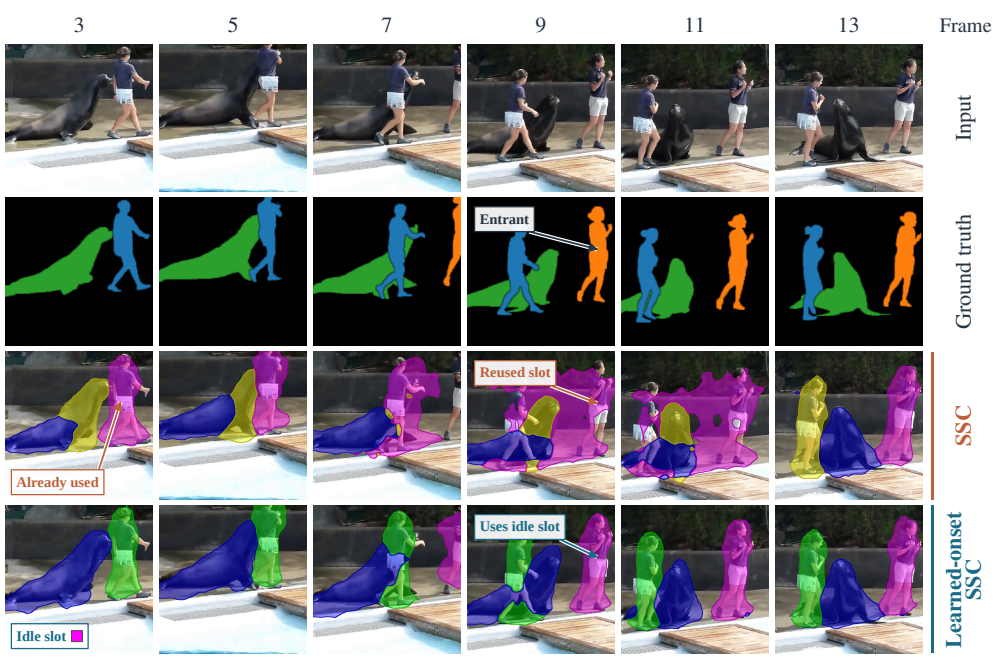


(a) An entrant appears at the image boundary

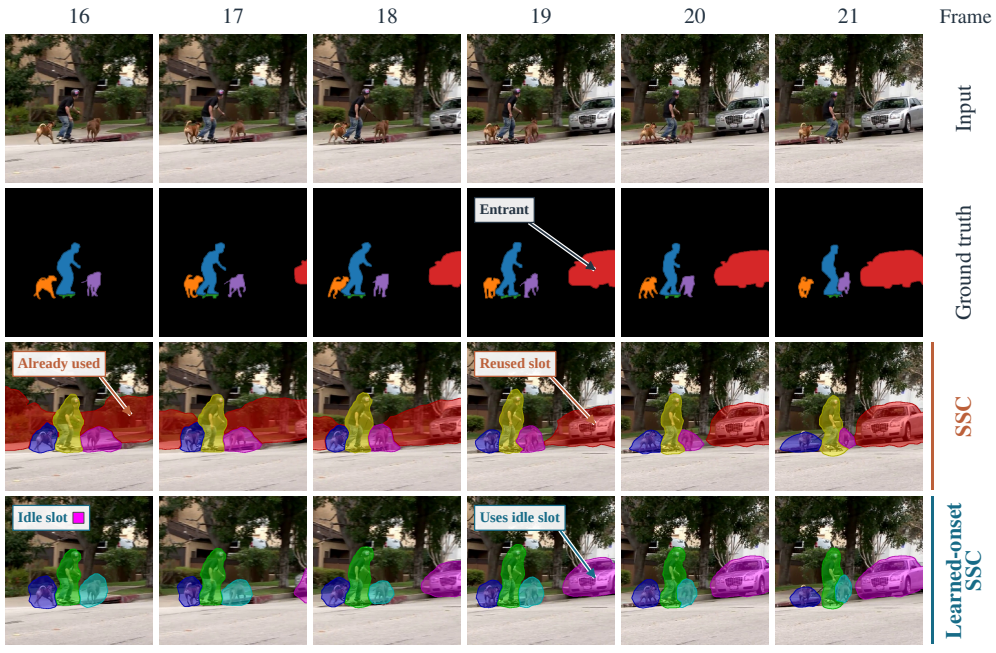


(b) Multiple objects enter over time

Figure 5: **Additional qualitative examples on MOVi-C/E.** Only the target entering objects are shown in the ground truth and predictions. Colors identify slots within each model and do not indicate correspondence across models.



(a) A second person enters the scene



(b) A vehicle enters the scene

Figure 6: **Additional qualitative examples on YouTube-VIS 2021.** All matched foreground objects are shown. Colors identify slots within each model and do not indicate correspondence across models.