IDENTITY WITHOUT PROPAGATION OR MATCHING: PARALLEL OBJECT-CENTRIC VIDEO LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video object-centric learning decomposes a video into a small set of object slots. The goal is that each slot binds one object and keeps that binding in every frame. The methods that lead the video benchmarks secure that binding by propagation, initializing each frame from the previous frame's slots through a learned transition model. We present CoSlot, a fully parallel video object-centric model that binds an entire clip in a joint pass, one slot-attention competition over all of its patches, and then refines every frame independently with the same module and no explicit propagation mechanism. It is trained with weak supervision: a differentiable agreement term in the loss compares the slot assignment, pooled over the whole clip, with imperfect reference masks from an unsupervised model or a foundation segmenter; no ground-truth mask is used anywhere. The identity it obtains is competitive: against three published baselines, CoSlot leads on clip-level identity on two of three benchmarks, by .496 to .443 Clip FG-ARI on YouTube-VIS, and matches the strongest on the third, at comparable mask quality. We provide a theory that explains this empirical performance. The output of the joint pass defines, for each clip, K convex cones in the projected feature space, one per slot and the same in every frame; an object whose patches stay in one cone keeps one slot in every frame, because every frame is partitioned by the same cones and no information passes between frames. Empirically, the cones the trained joint pass produces predict which slot owns an object significantly better than those of an untrained model.

1 INTRODUCTION

Object-centric learning decomposes a scene into a small set of representations that distinguish its objects, one per object. On video the same object must be assigned the same representation in every frame of the sequence, and that is the central difficulty: a decomposition computed frame by frame carries no correspondence between frames, since its representations are exchangeable and the one an object receives is arbitrary from one frame to the next, while the object itself changes pose, scale and visibility. Slot attention (Locatello et al., 2020) is the de facto module for extracting such a representation from a frame: a fixed number of slot vectors compete, through iterated attention over the frame's features, for what each will explain, and each comes to represent one object. Two threads of work bind identity across frames. The first, which leads the video benchmarks, uses one ingredient, *recurrence*: the previous frame's slots initialize the next frame's binding through a learned transition model (Kipf et al., 2022; Singh et al., 2022; Zadaianchuk et al., 2023; Manasyan et al., 2025; Seong et al., 2026; Moon et al., 2026), in training and at inference alike. A second thread of work binds frames independently and needs an alignment step instead: matching slot sets across time (Singh et al., 2024; Li et al., 2026), or full attention over every frame and position at a cost quadratic in clip length (Kabra et al., 2021).

In this paper, we address video object-centric learning with a simple architecture that requires neither recurrence nor an alignment step at inference. We introduce **CoSlot**, which binds an entire clip in a joint pass, one slot-attention competition over all of its patches, so that each slot comes to denote one object for the whole sequence; its slots are broadcast to every frame to seed a per-frame refinement by the same slot-attention weights, and a per-position decoder emits the masks (Figure 1). Slot attention on each frame runs independently of the others and in parallel. Identity is elicited by

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.

the training signal rather than propagated at inference. To train it, we introduce a differentiable, matching-free agreement term computed on a contingency table pooled over the frames used in training. It is a soft relative of the identity metric, compared with imperfect reference masks rather than ground truth — weak supervision — and applied under wide-baseline frame sampling. No human annotation is used anywhere, and only the model runs at inference. The reference masks are the strongest baseline's own released output on MOVi; on real video they come from a promptable foundation segmenter, SAM 3 (Carion et al., 2026), prompted with category names; we read no ground-truth mask.

This paper makes three contributions. First, a video object-centric model with no propagation and no alignment step that leads on clip-level identity on MOVi-C and YouTube-VIS, .813 to .796 and .496 to .443 Clip FG-ARI against the strongest baseline, and matches it on MOVi-E. Second, the training that makes this possible on one slot-attention module run at two scopes: an agreement term in the loss, pooled over the clip and computed against imperfect reference masks, weak supervision with no human annotation; a background penalty that keeps object slots off the background; and wide-baseline frame sampling. Third, a theory of where that identity comes from: the output of the joint pass defines, for each clip, K convex cones in the projected feature space, the same in every frame, so an object whose patches stay in one cone keeps one slot in every frame with nothing carried between frames, and two objects that no hyperplane through the origin separates must share a slot; on the trained models that partition predicts which slot owns an object far above chance, against near chance for the same architecture untrained.

2 Related work

The assignment of positions to objects is made by slot attention, an iterated competition of a fixed number of slots over the positions of a frame, each position going to the slot that wins its softmax (Greff et al., 2019; Locatello et al., 2020). On images, DINOSAUR replaced pixel reconstruction with the frozen patch features of a self-supervised transformer, which already cluster by object (Seitzer et al., 2023; Oquab et al., 2024).

On video the same object must be assigned the same slot in every frame, and the methods that lead the benchmarks obtain this by recurrence, in training and at inference alike. SAVi established the pattern, feeding the previous frame's slots in as the next frame's initialization through a learned transition model (Kipf et al., 2022), and later work kept that recurrence and changed what surrounds it: STEVE scaled the decoder (Singh et al., 2022), VideoSAUR predicted temporal feature similarities (Zadaianchuk et al., 2023), SOLV bound slots over local temporal windows (Aydemir et al., 2023), SlotContrast contrasted a slot against itself across adjacent frames (Manasyan et al., 2025), and SRL refines the representation itself (Seong et al., 2026). The strongest published Clip FG-ARI comes from SlotCurri (Moon et al., 2026), which adds a reconstruction-guided slot-count curriculum, a spatio-temporal structure loss, and a test-time cyclic inference sweep; the most recent work refines the recurrence rather than removing it (Zhao et al., 2026a;b).

Methods that do not recur need an alignment step instead. PSB binds every time step in parallel but keeps a separate slot set per time step, so slots must still be matched across time (Singh et al., 2024); SIMONe infers object and frame latents with a transformer over space and time, at a cost quadratic in clip length (Kabra et al., 2021). Closest to our premise, Grounded Correspondence (Li et al., 2026) argues that frozen self-supervised features already discriminate instances, so a learned predictor is unnecessary, and replaces it with Hungarian matching of each frame's slots to the previous frame's; identity is still a correspondence solved frame to frame, in sequence along the clip. CoSlot binds the clip jointly, with neither a matcher nor the quadratic cost, through slot attention's query–key asymmetry: one competition of K slots against all the patches of a clip has a cost linear in the clip length, because K is small and fixed.

Foundation segmenters trained on large mask corpora (Ravi et al., 2025; Carion et al., 2026) can provide a training signal for any video without annotating it: prompted with category names, they return object masks. Our training signal for identity is an agreement term against such reference masks, weak supervision that uses no human annotation. Training one model on another's outputs is distillation, whose gains over the source are well documented (Cho & Hariharan, 2019; Busbridge et al., 2025).

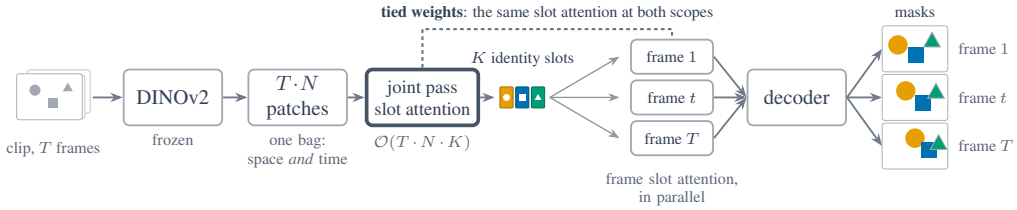


Figure 1: CoSlot. The joint pass, one slot-attention competition over all $T \cdot N$ patches of a clip, returns K clip-level slots; they are broadcast to every frame and refined by the same tied-weight module.

3 METHOD

3.1 ARCHITECTURE

Slot attention (Locatello et al., 2020) is the key architecture for learning object representations from frames and clips. It binds a set of input features to K slots. The slots start from learned initial vectors; in each iteration every input position is assigned to the slots by a softmax over slots, so the slots compete for each position, and each slot is then updated with the attention-weighted mean of the positions it won, through a GRU and a small MLP. The competition and update repeat a fixed number of times, three at each scope here, and the final attention is each slot's mask.

The architecture of CoSlot is presented in Figure 1. It runs slot attention twice, at two scopes. In the first pass, the joint pass, the frozen encoder's patches from all T frames are pooled into one set and K slots compete over all of them at once. In the second, per-frame pass, its K output slots are broadcast to every frame as the initialization of the same slot attention run on each frame independently, and a per-position decoder emits one mask per slot per frame.

Features. Let a clip be $x \in \mathbb{R}^{T\times 3\times H\times W}$. The frozen DINOv2 encoder yields $N = g^2$ patch features per frame on a $g \times g$ grid, one per 14×14-pixel patch of the $H \times W$ input, so $g = H/14$, and a small MLP projects them to D, giving $z \in \mathbb{R}^{T\times N\times D}$; everything below lives on this patch grid.

Joint pass. We flatten space and time together into one bag of $T \cdot N$ patches and run iterated slot attention with K slots over it once:

$$s,\ A\ =\ \mathrm{SlotAttention}\big(\mathrm{flatten}(z,\ T{\cdot}N,\ D)\big), \qquad s \in \mathbb{R}^{K\times D}, \quad A \in \mathbb{R}^{K\times T\cdot N}. \tag{1}$$

The softmax is over slots, so every patch in the clip is contested by all K slots at once, and the K slots returned each denote one object for the entire clip. The pass is affordable because slot attention is asymmetric, K queries against $T \cdot N$ keys, so it costs $\mathcal{O}(T \cdot N \cdot K)$, linear in T. It is also the sole source of cross-frame identity: every per-frame slot set below descends from it. Its slots start from K learned vectors, the same for every clip.

Per-frame pass. The K slots s are broadcast unchanged to every frame as the initialization of the same slot attention, run on each frame's own patches with the frame axis in the batch,

$$s_t,\ A_t\ =\ \mathrm{SlotAttention}\big(z_t;\ \mathrm{init} = s\big), \qquad t = 1, \ldots, T, \tag{2}$$

so the frames are processed independently and in parallel. The weights are tied so that the operator that formed the clip-level partition is the one that refines within it; under that constraint a per-frame mask can only sharpen the region its slot already owns. A per-position decoder then emits content and mask from a single network.

Inference. A clip is encoded once, the joint pass runs once, the frames are refined concurrently and the decoder emits the masks; there is no rollout and no test-time sweep.

3.2 OBJECTIVE

Following SlotCurri and SlotContrast, we optimize reconstruction in the frozen DINOv2 feature space (Seitzer et al., 2023), together with SlotCurri's SSIM term on the reconstructed feature volume, which favors spatially and temporally coherent reconstructions (Moon et al., 2026), and SlotContrast's self-supervised slot-contrastive loss, which pulls each slot toward itself in the adjacent

frame and away from the other slots (Manasyan et al., 2025). To these we add the agreement term that elicits identity from the slots, and a background penalty; both are defined below.

3.2.1 REFERENCE MASKS AND THE AGREEMENT TERM

A mask is a soft assignment of every patch of a frame to one of a set of segments, a number in $[0, 1]$ per patch and segment that sums to one over the segments. The output of the model is a mask over its slots, $a_{k,t,n}$, assigning patch n of frame t to slot k. The reference mask is fixed and supplied as weak supervision, and it is binary: $g_{j,t,n} \in \{0, 1\}$ assigns each patch to at most one of the J objects in the clip. A patch is *claimed* when it is assigned to some object and *unclaimed* otherwise; the split divides every frame into the foreground the agreement term acts on and the background the penalty below acts on. We form a single $K \times J$ contingency table for the whole clip by pooling over every frame and patch,

$$n_{kj} \;=\; \sum_{t=1}^{T} \sum_{n=1}^{N} a_{k,t,n}\, g_{j,t,n}, \qquad \mathcal{L}_{\text{agree}} \;=\; 1 - \mathrm{ARI}\big(n\big), \tag{3}$$

and take one minus the adjusted Rand index of that table evaluated on the real-valued n_{kj}. Its agreement term $A = \sum_{kj} \binom{n_{kj}}{2}$ counts the patch pairs our partition and the reference partition both place together. The term is differentiable, because the assignment is a softmax rather than an $\arg\max$; matching-free and permutation-invariant, since relabeling our slots permutes the rows of n and leaves the index unchanged; and per-clip by construction, since the pooling happens before the index is formed, which is what lets it see cross-frame structure at all. Concretely, a slot that owns object j in the first half of the clip and object j' in the second contributes a row of n split between two columns, whereas a per-frame index would see a pure row in every frame; the pooled index charges the switch and the per-frame one does not. The same holds when two slots divide one object over time: its column is split between two rows. Nothing in the term names a frame, so it is indifferent to which frames were sampled and to their order. It is applied to the decoded masks, the quantity that is scored, and to the joint attention, where identity is decided, and it is held off while only two slots are active and ramped in after each growth step, since a two-slot foreground–background split carries no object identity to agree with.

3.2.2 BACKGROUND PENALTY

The agreement term is silent on the background, because the reference mask assigns nothing there, yet the decoder's softmax runs over slots, so every background patch must be allocated to some slot; left to reconstruction alone, background is smeared across the object slots, which corrupts their masks and lets a slot bleed from one object toward another. With $f_k = \sum_{t,n} \sum_j g_{j,t,n}\, a_{k,t,n}$ slot k's mass on the claimed patches and $b_k = \sum_{t,n} \big(1 - \sum_j g_{j,t,n}\big)\, a_{k,t,n}$ its mass on the unclaimed ones,

$$\mathcal{L}_{\text{bg}} \;=\; \frac{1}{|\mathcal{B}|} \sum_{k} \frac{f_k}{\sum_{k'} f_{k'}}\; b_k, \tag{4}$$

with $|\mathcal{B}|$ the number of unclaimed patches and the foreground shares treated as constants. It charges a slot for the background it absorbs in proportion to the foreground it owns, so a slot that owns no object may take the background freely while object slots are pushed to stay on their objects: it is what allocates background to dedicated slots rather than smearing it across object slots.

3.3 TRAINING MECHANISMS

3.3.1 WIDE-BASELINE FRAME SAMPLING

A training example is eight frames drawn as a sorted random subset of the whole clip, redrawn every time the clip is visited. Temporally adjacent frames are near-duplicates, and a model shown them can satisfy a consistency objective by interpolating local changes between neighbors; wide gaps remove that shortcut, so the agreement term can only be satisfied by binding each object's patches across large changes in pose, scale, position and occlusion, on what the frozen features carry rather than on frame-to-frame continuity (§4.4).

3.3.2 SLOT-COUNT CURRICULUM

We employ SlotCurri's slot-count curriculum (Moon et al., 2026): training begins with two slots and the count grows in two steps to its final value, $2 \to 5 \to 11$ on MOVi-C, $2 \to 7 \to 15$ on MOVi-E and $2 \to 3 \to 7$ on YTVIS, at fixed points of the schedule. Beginning with few slots forces a coarse decomposition first, and slots are added where the remaining reconstruction error says they are needed: at each growth step the new initial slot vectors are spawned from the existing ones, allocated to parents in proportion to each slot's share of the reconstruction error, so the slot that explains its patches worst is split most, and each child is its parent displaced along a random direction by a length proportional to the distance to the parent's nearest neighboring vector, so that children start apart from the vectors already present. The agreement term is held off while two slots are active.

4 EXPERIMENTS

4.1 DATASETS, METRICS, AND PROTOCOL

We evaluate on MOVi-C and MOVi-E (Greff et al., 2022), synthetic multi-object video with ground-truth instance masks, and on YouTube-VIS 2021 (YTVIS) (Yang et al., 2019), real video of 40 object categories in clips of 12 to 62 frames. More details about the training and evaluation setup are in Appendix G.

We report three quantities. *Frame FG-ARI* is the adjusted Rand index over foreground pixels, computed per frame. It is blind to which slot holds an object from frame to frame. *Clip FG-ARI* pools one contingency table over the whole clip first, so a slot that changes object is penalized. It is the metric that requires cross-frame identity. *mBO* is the best-overlap IoU (intersection over union) per ground-truth mask. It measures boundaries rather than identity and responds strongly to how a model allocates background.

4.2 THE REFERENCE MASKS

The agreement term of §3.2 requires a reference mask for each training clip. On MOVi we take our reference masks from the released SlotCurri checkpoint, run under its own published cyclic protocol, its masks cached once and used as fixed targets. These targets are imperfect (Clip FG-ARI .80 on MOVi-C) and are produced by the model we compare against. Because the supervisory signal is the baseline itself, a margin over SlotCurri cannot be attributed to a stronger training signal, and on MOVi-C the model exceeds the source of its own supervision, so the formulation extracts more structure than it was shown. A model trained on another model's outputs can exceed them through the regularizing effect of distillation alone (Cho & Hariharan, 2019; Müller et al., 2019; Zhou et al., 2021; Busbridge et al., 2025), so we treat the accuracy gain as supporting evidence rather than as the contribution.

On YouTube-VIS the reference mask must likewise come from a model, and using SlotCurri again fails on real video for a measurable reason: its per-frame masks collapse on natural video. On the same clips, SlotCurri's Frame FG-ARI falls from .83 on MOVi-C to .49 on YTVIS (Table 1). The per-frame quality of the reference bounds what the agreement term can teach: trained on SlotCurri's YTVIS masks, CoSlot plateaus at .37 Clip FG-ARI, below SlotCurri's own .44. Real video requires a real per-frame segmenter. We therefore use SAM 3 (Carion et al., 2026), a promptable foundation segmenter whose knowledge of objects comes from outside this architecture and this dataset, prompted with the dataset's category vocabulary and cached once. Scored as a model, its masks are far more accurate per frame than SlotCurri's (Appendix F). And it supplies no background: it returns object tracks rather than a partition, so our model receives no signal there.

4.3 RESULTS

Table 1 scores every method on the same clips; for YTVIS we follow SlotCurri's validation protocol, its 300 held-out training clips. We compare against the three strongest video object-centric models: SlotContrast (Manasyan et al., 2025), SRL (Seong et al., 2026) and SlotCurri (Moon et al., 2026), the last being the released checkpoint whose MOVi masks serve as our reference masks. Our numbers

Table 1: All three benchmarks, one standard evaluation.

	MOVi-C			MOVi-E			YTVIS		
method	Frame FG-ARI	Clip FG-ARI	mBO	Frame FG-ARI	Clip FG-ARI	mBO	Frame FG-ARI	Clip FG-ARI	mBO
SlotContrast	.7660	.7014	.3247	.8409	.8182	**.2971**	.4549	.3630	.3285
SRL	.8019	.7410	.3422	—	—	—	.4813	.4345	.3658
SlotCurri	.8278	.7958	.3281	.8449	.8308	.2920	.4887	.4426	.3449
CoSlot (ours)	**.8412** ±.0015	**.8125** ±.0010	**.3430** ±.0089	**.8457** ±.0028	**.8322** ±.0030	.2929 ±.0018	**.5687** ±.0062	**.4961** ±.0023	**.3709** ±.0025

±: sd over three seeds, each the mean of its last three validations.

are the mean over three seeds, each the mean of its last three validations; each baseline is its single released checkpoint.

On the main metric for object identity across a clip, Clip FG-ARI, CoSlot leads every baseline on YTVIS by the largest margin in the table, well outside our seed deviation. On MOVi-C it out-scores every baseline. On MOVi-E it is level with SlotCurri, within one seed deviation, and ahead of the remaining baselines.

Figure 2 reads the YTVIS lead clip by clip. CoSlot is above SlotCurri on 151 of the 300 clips, within .01 on 76 and below on 73 (sign test over the discordant clips, $p < 10^{-6}$), and the ties are not close contests: 63 of the 76 are clips on which both models score at or below .1, and on 51 both score exactly zero. Where the two differ they differ in our favor by more, 52 clips won by over .2 against 30 lost. The lead holds in every length band from 12 to 62 frames, and neither model's score depends on length (correlation $+.08$ for CoSlot, $+.11$ for SlotCurri), although CoSlot was trained on eight-frame samples.

Frame FG-ARI shows that CoSlot out-scores every baseline on MOVi-C and YTVIS and is level with SlotCurri on MOVi-E, so the clip-level consistency is not bought by sacrificing per-frame masks. On YTVIS the Frame FG-ARI gain over SlotCurri is larger than the Clip FG-ARI gain, so most of that lead is per-frame mask quality inherited from the reference masks and carried through the clip by the joint pass. MOVi tests the same contribution independently of the source, since there the source is the baseline itself and the model exceeds it.

On mBO, CoSlot is level with the best method on MOVi-C and on YTVIS, within seed deviation, and trails SlotContrast on MOVi-E by a small margin. mBO takes the best overlap for every ground-truth object, so on these datasets it is dominated by where the background goes: an object slot that absorbs surrounding background loses overlap with its identity intact. Our agreement term shapes the foreground only, so background assignment is left to the feature reconstruction and the background penalty.

4.4 Ablation Study

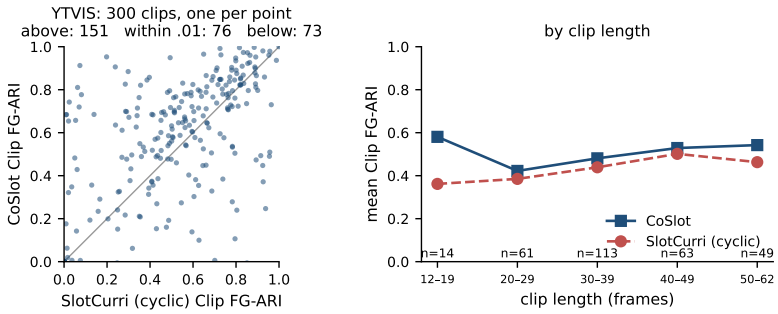


Figure 2: YTVIS validation. Left: Clip FG-ARI of CoSlot against SlotCurri, one point per clip; points above the diagonal are clips CoSlot wins. Right: mean Clip FG-ARI by clip length, with the clip count per band.

	MOVi-C		YTVIS	
removed	Δ Clip FG-ARI	Δ mBO	Δ Clip FG-ARI	Δ mBO
agreement term	−.1784	−.0713	−.3421	−.2249
contiguous frames	−.0551	−.1032	+.0037	+.0118
untied per-frame attention	+.0004	−.0130	−.0152	−.0063
slot-contrastive term	−.0048	−.0088	+.0181	+.0049
background penalty	—	—	+.0492	−.0709

Table 2: Ablation. Change in Clip FG-ARI and mBO when one key is removed (*contiguous frames* replaces wide-baseline sampling). Absolute scores in Table 6.

Table 2 removes one ingredient of §3 at a time, the agreement term, the wide-baseline sampling, the tie between the two slot-attention passes, the slot-contrastive term and the background penalty; every figure is a mean over the final three validations, read after the curriculum's final slot-growth step.

Removing the agreement term is by far the largest degradation on both datasets, costly on MOVi-C and disabling on YTVIS. Wide-baseline sampling is load-bearing on MOVi-C, where training on consecutive frames makes the representation substantially worse, so temporal identity is not an adjacency artifact; on YTVIS the same change is on par. The structural ablation follows the same pattern: untying the per-frame slot attention from the joint pass, which strictly increases capacity, buys nothing on MOVi-C and costs Clip FG-ARI on YTVIS, so the tie is doing work and added capacity is not what produces the result. No other component is a source of identity: removing the inherited slot-contrastive term is a mild loss on MOVi-C and a gain of .018 Clip FG-ARI on YTVIS; we keep it because it steadies the background slots across frames (Figure 3). The background penalty is a dial that trades mBO for Clip FG-ARI.

5 A THEORY OF WHERE IDENTITY COMES FROM

Nothing in CoSlot carries state between frames or reads a frame index, so its identity mechanism can be stated exactly, as facts about the model itself. Two such facts are derived below: what the per-frame pass's first attention iteration, called *iteration-0* below, can express (§5.1), and what the objective prefers (§5.2); each yields predictions that could have failed, and for each we state the prediction, how it was measured, and what was observed. The first of these predictions, that the joint pass would have to fix identity and the per-frame pass would refine it, was stated before any trained model was probed. The others follow from the derivation and were measured after it.

The account below takes the joint pass's output as given. Any module that returns K vectors could seed the per-frame pass; the joint pass is one such module, slot attention run over the clip's projected patches, that is, a soft clustering of them. The K slots it returns define K convex cones in the projected feature space (§5.1), and the account says which objects keep one slot across the clip and which pairs of objects no placement of the cones can separate. It says nothing about how the joint pass computes the cones; that is measured. Two kinds of failure are therefore kept apart. The theory forbids one: two objects that no hyperplane through the origin separates share a slot, at every parameter setting. It does not forbid the other. An object whose patches could have been given a cone of their own, but were not, is a failure of the joint pass. How often that happens is measured below; the theory does not predict it.

All measurements use the trained models of Table 1, three seeds per dataset. The first prediction is measured on every seed and reported as the mean; the others on one seed per dataset. Its control is the same architecture with the same K and random weights, the median of three draws. An object's *owner* in a frame is the slot to which the decoded mask assigns most of its patches, and it has *one owner* when that slot is the same in every frame it is visible in. Its *iteration-0 slot* in a frame is the slot the per-frame pass's raw first-iteration logits prefer for most of its patches, read before any refinement, and it is *stable* when that slot is the same in every frame. An object is *merged* when its owning slot also owns at least a fifth of another object's mass. Ground truth delimits objects and nothing else; no reference mask enters any measurement.

5.1 THE FIRST ITERATION FIXES A PARTITION OF FEATURE SPACE

The per-frame pass is seeded by the joint pass's K slots, broadcast unchanged to every frame. Its first attention iteration, iteration-0, scores every patch against these seeds before any update, and that

score is linear in the patch: for a patch $\tilde{z} \in \mathbb{R}^D$ after the learned projection and input normalization,

$$a_k(\tilde{z}) \;=\; \langle u_k,\, \tilde{z}\rangle, \qquad u_k \;=\; \tfrac{1}{\sqrt{D}}\, W_\kappa^\top q(s_k), \tag{5}$$

so slot k enters the competition through a single direction u_k, and the slot each patch prefers, $\arg\max_k a_k$, partitions $\mathbb{R}^D$ into K convex cones $\mathcal{C}_k = \{\tilde{z} : \langle u_k - u_{k'}, \tilde{z}\rangle > 0 \;\forall k' \neq k\}$, fixed for the whole clip by the seeds and the projections alone. Let object j occupy the patch set $Z_j \subset \mathbb{R}^D$, pooled over the frames in which it appears. Iteration-0 is the one point where the partition is shared: later iterations update each slot from the patches it won in that frame, so their partition is frame-dependent. The propositions below, proved in Appendix B, constrain the assignment at iteration-0, not the slot vectors and not the decoded mask; what carries from that assignment to the final owner is measured, not assumed.

Proposition 1 (Frame-invariant assignment)**.** *Object j is assigned to one slot k at iteration-0 in every frame it occupies if and only if $Z_j \subset \mathcal{C}_k$. No information passes between frames.*

Proposition 2 (Expressibility)**.** *There exist parameters under which every object in a clip is assigned its own slot at iteration-0 in every frame if and only if $Z_1, \ldots, Z_J$ are multiclass linearly separable through the origin: there is an injection $j \mapsto k(j)$ and directions $u_1, \ldots, u_K$ such that $\langle u_{k(j)} - u_{k'}, z\rangle > 0$ for all $z \in Z_j$ and $k' \neq k(j)$. In particular, two objects whose patch sets are not separated by a hyperplane through the origin share a cell at* every *parameter setting.*

From these propositions we derive two properties the trained models should have, and check them.

An object whose iteration-0 slot is stable keeps one owner, an object whose iteration-0 slot changes does not, and only the trained partition has this property. Proposition 1 makes stability sufficient for a frame-invariant assignment at iteration-0; if that assignment decides the final owner, stability should predict one owner far better than its absence, and the control is needed because a random partition also has cells, so stability alone would predict nothing. Splitting the objects of every validation clip by whether their iteration-0 slot is stable, a stable object keeps one owner 5.2 and 8.0 times as often as a changing one on MOVi-C and YTVIS; with random weights the ratio is 1.3 and 1.2, near one (Table 3, block 1). The claim is that stability decides identity, not that the seeds fix every object: only 35% and 47% of objects are stable at iteration-0.

	MOVi-C	YTVIS
(1) how many times more often an object keeps one owner when its iteration-0 slot is stable than when it changes		
trained	5.2×	8.0×
untrained	1.3×	1.2×
(2) fraction of an object's patches that a hyperplane through the origin separates in the trained projection, mean over objects		
merged objects	.949	.919
kept-apart objects	.984	.971
(3) single-switch objects whose split lies in the central band $[.4, .6)$; a uniform distribution would give 20%		
observed	6%	11%

Table 3: Three predictions of the account, on the trained models. *untrained*: same architecture and K, random weights, median of three draws.

The objects that remain merged are the ones the projection separates least. For each validation clip with at least two objects we fit a homogeneous multiclass linear classifier to the foreground patches labeled by ground-truth object, in the trained projection (Appendix C); its training accuracy is the fraction of patches separable, joined per object with whether the model merged that object. The objects the model merges are less separable than the ones it keeps apart: $.949$ against $.984$ on MOVi-C and $.919$ against $.971$ on YTVIS (Table 3, block 2), with rank-AUROC $.66$ on both datasets (Table 5). Where CoSlot fails, it fails where the proposition says it must.

5.2 The objective prefers one slot per object for the whole clip

The agreement term $A = \sum_{kj} \binom{n_{kj}}{2}$ is convex in each entry, so at fixed marginals the index is maximal exactly when every slot is pure and every object unsplit, and because the table is pooled over frames that maximizer is one slot per object *for the whole clip*. The shortfall has a closed form, one slot at a time: with $m_k = \sum_j n_{kj}$ and $p_{kj} = n_{kj}/m_k$,

$$\tbinom{m_k}{2} - \sum_j \tbinom{n_{kj}}{2} \;=\; \tfrac{1}{2}\Big(m_k^2 - \sum_j n_{kj}^2\Big) \;=\; \tfrac{1}{2}\, m_k^2\, \mathrm{Gini}(p_k), \tag{6}$$

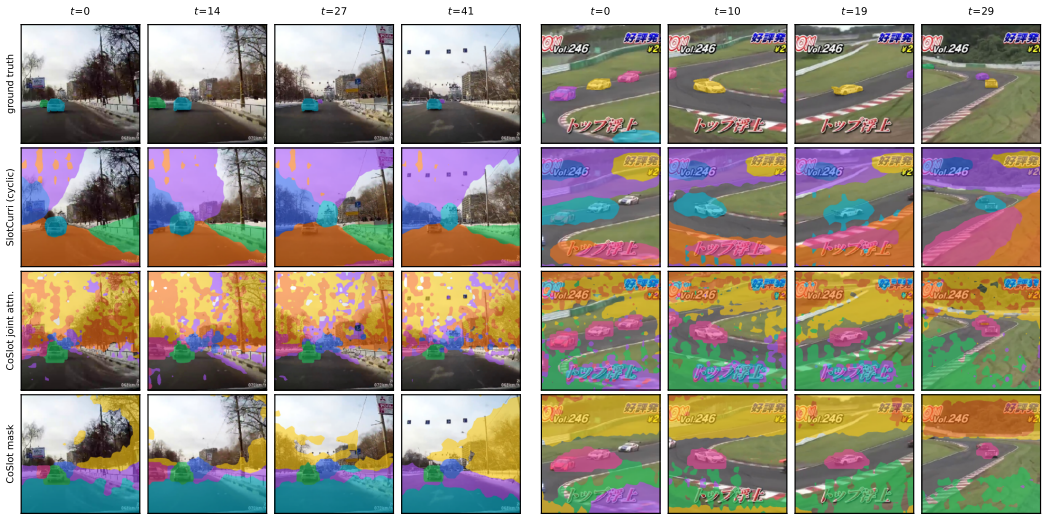


Figure 3: Two YTVIS clips with at least three objects: our largest margin over SlotCurri (left) and our median loss (right). Rows: ground truth, SlotCurri, our joint attention before refinement, our mask; color is slot index.

a Gini impurity weighted by squared mass, and symmetrically per object. The weighting is visible in the trained models: only 37% and 39% of objects keep one owner for their whole visible life on MOVi-C and YTVIS, yet over within-object patch pairs identity holds for 96% and 89%, because the switches sit where (6) prices them lowest, on small objects and brief excursions.

Residual splits are lopsided, not balanced. A slot split between two objects in proportion α costs $m_k^2\alpha(1-\alpha)$, maximal at an even split, so even splits should be removed first and the ones that survive should sit near $\alpha = 0$ or 1. Among objects with exactly one owner switch we record the share of the object's mass held by its first owner. Under a uniform distribution 20% of these shares would fall in the central band $[.4, .6)$, that is, near an even split; 6% and 11% do (Table 3, block 3), and 71% and 43% lie within $.1$ of an end (Table 5).

6 Limitations and Conclusion

Video object-centric models obtain identity across frames in two ways, by making each frame depend on the last or by an alignment step across frames. CoSlot is a third route with neither: one joint pass binds the clip, the same slot attention refines each frame independently, and the result is competitive with the strongest models of both. Its identity is elicited rather than carried: a clip-pooled agreement term acts on the per-frame pass's first attention iteration, whose partition of feature space, fixed for the whole clip, predicts slot ownership before any refinement, and the objects that continue to merge are the ones that partition separates least. The formulation requires weak supervision, a reference mask for every training clip.

The theory has two limits. First, it holds at iteration-0 of the per-frame pass. Later iterations update each slot from the frame's own patches, so their partition is no longer shared across frames. What carries from the iteration-0 assignment to the final owner is therefore measured, not derived. Second, it says nothing specific to the joint pass. It takes the K slots as given. It does not explain how slot attention over the clip's patches arrives at cones that separate the objects, and it does not explain why it falls short when it does.

These limits point to three directions for future work. The first is an architecture, or a training signal, that leads the joint pass to separating cones more often than slot attention over the clip's patches does now. The second is a projection of the frozen features chosen for that separability, rather than inherited from the reconstruction loss. The third is a self- or semi-supervised replacement for the reference masks, for which propagation across frames, used as a training signal rather than at inference, is a natural source.

AI USE STATEMENT

We used generative AI tools, a coding and writing assistant, throughout this work, in five ways. They drafted and polished text, including sections of this manuscript that we then revised. They searched for and summarized related work, which we checked against the original sources and cite only when refereed. They wrote, refactored and ran the training, evaluation, probe and figure-generation code and summarized its output. They proposed and refined analyses and controls in discussion with us, several of which became the measurements of §5 and the appendices. And they wrote out the derivations and proofs of Appendix B, which we checked line by line against the implementation. They did not generate experimental results or numbers: every quantitative claim traces to a file produced by a measurement in our own codebase, and the figures that carry measurements are generated from those files by scripts rather than transcribed. Model code and metric implementations were verified against behavior-preserving gates that require bit-identical outputs, and baseline figures were reproduced from released checkpoints through our own evaluation rather than quoted. Several claims in earlier drafts were retracted when measurement contradicted them. We reviewed all AI-assisted work and take responsibility for the final content, including text, claims, proofs and artifacts produced with its aid.

REPRODUCIBILITY STATEMENT

The scoring protocol is specified in §G and is the single path by which every number in this paper — ours and every baseline's — was produced; baselines are re-scored from their publicly released checkpoints rather than quoted, and SlotCurri is run under its own published cyclic protocol. Dataset geometry, clip counts, slot counts and evaluation length are given in §4. The reference masks used for supervision are the released baseline's own output on MOVi and SAM 3's on YTVIS, cached once, so they are reconstructible by anyone with those checkpoints. The architecture is specified in §3, including the tied-weight relationship between the two slot-attention scopes and the fact that the clip-level identity is broadcast unchanged to every frame, which together determine the model's behavior at initialization. All reported metrics are means over the last several evaluations of a run rather than best-of values. Code, the training configurations, and the scripts that generate the measurement figures will be released; an anonymized copy is available to reviewers.

ETHICS STATEMENT

This work uses two synthetic, procedurally generated video datasets and one public real-video benchmark (YouTube-VIS 2021), used under its terms; it involves no human subjects, no personally identifying data, and no human annotation collected by us. Object-centric segmentation is a dual-use capability in the same sense as segmentation generally: the same representation that supports scene understanding could support surveillance applications. We see no risk specific to this work beyond that shared with the segmentation literature. Our models are trained against masks produced by released research checkpoints, whose license terms we observe.

REFERENCES

Görkay Aydemir, Weidi Xie, and Fatma Güney. Self-supervised object-centric learning for videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 32879–32899, 2023.

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russ Webb. Distillation scaling laws. In *International Conference on Machine Learning (ICML)*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations (ICLR)*, 2026.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

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 (ICML)*, pp. 2424–2433, 2019.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti Derek Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3749–3761, 2022.

Rishabh Kabra, Daniel Zoran, Goker Erdogan, Loic Matthey, Antonia Creswell, Matthew Botvinick, Alexander Lerchner, and Christopher P. Burgess. Simone: View-invariant, temporally-abstracted object representations via unsupervised video decomposition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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.

Zhiyuan Li, Rongzhen Zhao, Wenyan Yang, Wenshuai Zhao, Pekka Marttinen, and Joni Pajarinen. Rethinking temporal consistency in video object-centric learning: From prediction to correspondence. In *International Conference on Machine Learning (ICML)*, 2026.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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

WonJun Moon, Hyun Seok Seong, and Jae-Pil Heo. Reconstruction-guided slot curriculum: Addressing object over-fragmentation in video object-centric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025.

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 *International Conference on Learning Representations (ICLR)*, 2023.

Hyun Seok Seong, WonJun Moon, and Jae-Pil Heo. From vicious to virtuous cycles: Synergistic representation learning for unsupervised video object-centric learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Gautam Singh, Yue Wang, Jiawei Yang, Boris Ivanovic, Sungjin Ahn, Marco Pavone, and Tong Che. Parallelized spatiotemporal slot binding for videos. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5188–5197, 2019.

Andrii Zadaianchuk, Maximilian Seitzer, and Georg Martius. Object-centric learning for real-world videos by predicting temporal feature similarities. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Rongzhen Zhao, Jian Li, Juho Kannala, and Joni Pajarinen. Predicting video slot attention queries from random slot-feature pairs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026a.

Rongzhen Zhao, Wenyan Yang, Juho Kannala, and Joni Pajarinen. Smoothing slot attention iterations and recurrences. In *International Conference on Machine Learning (ICML)*, 2026b.

Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. In *International Conference on Learning Representations (ICLR)*, 2021.

APPENDIX CONTENTS

A Performance per seed 13
B Derivations behind §5 13
C Testing Proposition 2: are objects linearly separable at iteration-0? 14
D All mechanism measurements 15
E Qualitative results 16
F The YTVIS reference masks 17
G Training and evaluation details 19

A PERFORMANCE PER SEED

Table 1's $\pm$ is the standard deviation over seeds, each seed being the mean of its last three validations. With three seeds the deviation is itself noisy, so the seeds are listed below. The within-seed spread of the last three validations is at most 0.0034 on every dataset, below the across-seed spread, so a trailing mean summarizes a converged tail. On MOVi-C, Clip FG-ARI repeats to 0.0010 across seeds while mBO moves by 0.0089. Our $\pm$ is not comparable to the baselines': two of the three report no dispersion and the third reports one without defining what it varies over.

	Clip FG-ARI			mBO		
	seed 1	seed 2	seed 3	seed 1	seed 2	seed 3
MOVi-C	0.8134	0.8127	0.8115	0.3418	0.3348	0.3524
MOVi-E	0.8355	0.8313	0.8297	0.2924	0.2915	0.2949
YTVIS	0.4936	0.4969	0.4979	0.3717	0.3681	0.3728

B DERIVATIONS BEHIND §5

The objective and the derivations below use the adjusted Rand index in the following form. With $m_k = \sum_j n_{kj}$ a slot's mass, $c_j = \sum_k n_{kj}$ an object's, $M = \sum_{kj} n_{kj}$, $a. = \sum_k \binom{m_k}{2}$ and $b. = \sum_j \binom{c_j}{2}$,

$$\mathrm{ARI}(n) = \frac{A - \mathbb{E}}{\frac{1}{2}(a. + b.) - \mathbb{E}}, \qquad A = \sum_k \sum_j \binom{n_{kj}}{2}, \qquad \mathbb{E} = \frac{a.\, b.}{\binom{M}{2}}, \tag{7}$$

with $\binom{x}{2} = \frac{1}{2}x(x-1)$. $\mathbb{E}$ is A's value under independent partitions with the same marginals and the denominator is the largest A those marginals permit, so both depend on m_k and c_j alone — which is what the affine-in-A claim in §5 uses.

$A = \sum_{kj} \binom{n_{kj}}{2}$ sums pair counts that are convex in n_{kj}, so a slot of mass $m_k = \sum_j n_{kj}$ contributes at most $\binom{m_k}{2}$, with equality only when all of its mass sits on one object; at fixed marginals the adjusted index is affine and increasing in A. A is symmetric in k and j, so the same argument per object, with column mass $c_j = \sum_k n_{kj}$, bounds it by $\sum_j \binom{c_j}{2}$, with equality only when no object is split; a split object costs $\frac{1}{2}c_j^2$ Gini of its column, as a mixed slot costs (6). Equality in both requires the slot and object marginals to be matchable, so a clip with $J > K$ objects cannot attain the bound; MOVi-E has clips with $J = 19$ against $K = 15$.

Equation (6) is the shortfall in the agreement term, not the index itself, and it is exact, so it states how the cost of an identity error scales: quadratically in mass, and for a slot split between two objects in proportion α as $m_k^2\alpha(1-\alpha)$, maximal at $\alpha = \frac{1}{2}$ and vanishing at either extreme. The objective therefore pays most to remove an even split and least to finish a lopsided one. α is a fraction of mass, not of time; the two coincide only when an object's mass is uniform over the clip.

The per-frame pass is initialized by the joint pass's output, the same K vectors for every frame, and per-frame attention then runs independently, so slot k's assignment in frame t is a function of that frame alone. Slot attention's attend-and-aggregate step is a soft k-means assignment with learned projections (Locatello et al., 2020), followed by the GRU and MLP, so the broadcast vectors act as cluster seeds: a seed fixes which cluster it converges to, not its converged value. The directions u_k

in (5) are the image of K generic seeds under two unconstrained linear maps, so for $K \leq D$ every partition of that form is realizable.

Iteration-0 is the first attention, computed from the seed slots before the GRU update; we read the raw slot logits a_k, upstream of slot attention's per-slot renormalization, which changes the assignment on about a third of patches. The softmax over slots is monotone in a_k, so the slot preferred at iteration-0 is $\arg\max_k a_k$. The cells $\mathcal{C}_k$ are intersections of homogeneous half-spaces, so $Z_j \subset \mathcal{C}_k$ is the condition of Proposition 1, and Proposition 2 restates it for all objects at once; its two-object case is the existence of a separating homogeneous hyperplane, which is a cell boundary. Position cannot separate two look-alikes: the affinity reads projected features only and receives no position code. We verified (5) against the implementation by reconstructing the model's iteration-0 attention from u_k and $\tilde{z}$ alone, to a mean absolute deviation below 10^{-9} on both datasets (Table 5), and confirmed the broadcast seeds are frame-constant to 0. The separability condition of Proposition 2 is decidable per clip, and Appendix C tests it on all three datasets.

Proof of Proposition 1. Fix a frame t and a patch $\tilde{z}$ of object j in it. By (5) its iteration-0 slot is $\arg\max_k \langle u_k, \tilde{z} \rangle$, which equals k exactly when $\langle u_k - u_{k'}, \tilde{z} \rangle > 0$ for every $k' \neq k$, that is, when $\tilde{z} \in \mathcal{C}_k$; a patch on a cell boundary, a tie, is non-generic, belongs to no cell and is excluded. The directions u_k are computed from the broadcast seeds and are therefore the same in every frame, so object j is assigned slot k at every patch in every frame it occupies exactly when every $\tilde{z} \in Z_j$ lies in $\mathcal{C}_k$, that is, when $Z_j \subset \mathcal{C}_k$. No quantity indexed by t enters. □

Proof of Proposition 2. ($\Rightarrow$) If some parameters assign object j its own slot $k(j)$ at iteration-0 in every frame, then $j \mapsto k(j)$ is an injection and, by Proposition 1, $Z_j \subset \mathcal{C}_{k(j)}$, that is, $\langle u_{k(j)} - u_{k'}, z \rangle > 0$ for all $z \in Z_j$ and $k' \neq k(j)$: the directions $u_1, \ldots, u_K$ of those parameters separate $Z_1, \ldots, Z_J$ through the origin. ($\Leftarrow$) Given an injection and directions $u_1, \ldots, u_K$ with that property, parameters realizing them exist: $u_k = W_\kappa^\top q(s_k)/\sqrt{D}$ with q and W_κ linear and bias-free, so with $W_\kappa = I$ it suffices that q map the normalized seeds $\mathrm{LN}(s_k)$ to $\sqrt{D}\, u_k$, and for $K < D$ (here $K \leq 15$ against $D \geq 64$) and seeds whose normalized vectors are linearly independent, which holds for generic seeds, such a linear map exists. Proposition 1 then gives the assignment in every frame. The seeds are treated as free vectors here; in the model they are the joint pass's output, which the account takes as given. □

If the seeds are identical, all u_k coincide, every cell boundary degenerates, assignment is settled by tie-breaking, and no identity is expressible; slot-permutation symmetry must be broken before binding can begin. Both propositions constrain the assignment map rather than the slot vector, so identity is compatible with a seed being overwritten by refinement, which Table 5 shows it is on MOVi. They stop at iteration-0: extending them to the converged slots would require the iteration to be contractive on the induced partition, which we do not assume.

C Testing Proposition 2: are objects linearly separable at iteration-0?

Proposition 2 is an if-and-only-if: the model can give every object of a clip its own slot at iteration-0 exactly when the objects' projected patch sets are multiclass linearly separable through the origin. Both directions are decidable on frozen data without reading the trained slots, so we test the proposition directly.

This is a supervised probe with the ground truth in hand, and it measures what the feature space permits, not what the model does. For each validation clip we pool the patches of all of its frames, bring the ground-truth instance masks to the patch grid by nearest sampling, keep the foreground patches with their ground-truth object as the label, and draw up to 6,000 of them uniformly at random. On these we fit a multiclass linear support-vector classifier with no intercept, so that every decision boundary is a hyperplane through the origin as the cell boundaries of (5) are, with a large margin penalty ($C = 10^4$) so that the fit approaches the hard-margin solution. The quantity reported is its accuracy on the patches it was fit to: the fraction of foreground patches that some choice of K directions assigns to their own object, and 1.0 means the clip's ground-truth partition is exactly realizable at iteration-0. There is no held-out set and none is wanted, since the question is whether

Table 4: Linear separability at iteration-0 (Proposition 2), per clip: patches a homogeneous multiclass linear classifier separates, and clips it separates exactly. *raw*: frozen features; *rand*/*train*: the D-dim projection at random/trained weights. Last block: per-object separability in the trained space, merged vs. kept objects, with rank-AUROC.

			patches separable (%)			clips exact (%)			obj. sep., train		
dataset	D	K	raw	rand	train	raw	rand	train	merged	kept	AUROC
MOVi-C	64	11	100.0	97.0	98.4	99.6	17.6	29.6	0.949	0.984	0.66
MOVi-E	128	15	100.0	99.7	99.9	100.0	54.4	72.8	0.999	0.999	0.50
YTVIS	64	7	100.0	95.2	96.7	100.0	21.0	26.9	0.919	0.971	0.66

a partition exists, so the classifier is meant to fit as well as it can. Its accuracy is therefore a ceiling: the most any K cones, and so the model's own, can separate at iteration-0. A single fit with finite C can stop short of the true maximum, so the ceiling is estimated from below. Per object, the accuracy is taken over that object's own patches, and an object is merged as in §5, when its owning slot also owns at least a fifth of another object's mass. We fit in three feature spaces on the same patches: the frozen encoder features, the trained patch projection (the space (5) reads), and a freshly initialized projection of the same dimension D, as a control for what training adds. The raw space lives in the encoder's dimension (384 or 768), where linear separability is cheaper, so it bounds what any projection can preserve; the like-for-like comparison is random against trained at fixed D. Runs use the trained models of Table 1 on the full validation splits.

Table 4 reads in three parts, of which only the third is a prediction stated in §5. The first is the check that the account is not vacuous, since if the frozen features did not separate the objects no projection could and the account would predict failure everywhere; the second is a consequence of §5.2, that an objective which pays for every merge should move the projection toward separability, and it was measured after the derivation. The condition is satisfiable: in the frozen features 100.0% of foreground patches are separable, and every clip of MOVi-E and YTVIS and all but one of the 250 MOVi-C clips exactly. Training makes it hold in the model's own space: after projecting to D the condition survives on 96–100% of patches, and the trained projection is exactly realizable on more clips than a random projection of the same width (30% vs 18% on MOVi-C, 27% vs 21% on YTVIS, 73% vs 54% on MOVi-E), so training moves the projection toward separability. Where the condition fails, the model merges: the objects it fuses in its final masks are less separable in the trained space than the ones it keeps apart (0.949 vs 0.984 on MOVi-C, 0.919 vs 0.971 on YTVIS; rank-AUROC 0.66 and 0.66). MOVi-E is the control: its projection (D=128) separates 99.9% of objects either way, there is no merged-versus-kept gap (AUROC 0.50), and merge is not its binding failure.

Separability is necessary, not sufficient. The classifier chooses K directions for each clip with the ground-truth labels in hand; the model's directions are the joint pass's output for that clip, produced without labels and read through one projection shared by every clip. A separable clip can therefore still be bound wrongly, and only a non-separable one cannot be bound rightly. This is why exact separability of the trained projection on 30% and 27% of clips is consistent with 41% and 56% of objects merged (Table 5), and why the last block compares more against less separable objects rather than separable against not. Switches are the other proposition: an object changes owner when its patch set straddles a cell boundary, which pairwise separability does not exclude.

D ALL MECHANISM MEASUREMENTS

Table 5 is the complete record behind §5: every probe on the trained models, three datasets across, with standard deviations where a probe ran on more than one seed. Its blocks follow the numbering of the section. Blocks (1) to (3) are the three predictions of Table 3 with every quantity the text quotes, including the raw one-owner rates behind the ratios of block (1); the untrained control's raw rates vary from draw to draw, since a random partition can collapse onto one slot, while their ratio stays near one on every draw, which is why the ratio is the reported statistic. Block (1) also carries two measurements of the joint pass itself, one seed each. The first is the share of an object's patches that its majority iteration-0 cone holds, 87% on all three datasets, against the oracle ceiling of Appendix C on the same objects, 97%, 100% and 94%. The second is the stability rate on clips that are exactly separable in the trained projection against the rest, 48% against 33%, 62% against

Table 5: Every measurement behind §5 and Table 3, on the trained models. $\pm$: sd over seeds where a probe ran on more than one. *untrained*: same architecture and K, random weights, median of three draws. AUROC: probability that a kept object is more separable than a merged one.

	MOVi-C	MOVi-E	YTVIS
equation (5) *against the implementation*			
iteration-0 attention rebuilt from u_k and $\tilde{z}$: mean / max abs. deviation	9e-10 / 1e-7	–	3e-10 / 2e-8
broadcast seeds, max $\lvert\Delta\rvert$ across frames	0.0	0.0	0.0
(1) how many times more often an object keeps one owner when its iteration-0 slot is stable than when it changes			
one-owner rate (%), stable / changing	$78_{\pm 0}$ / $15_{\pm 0}$	$87_{\pm 0}$ / $20_{\pm 1}$	$75_{\pm 4}$ / $10_{\pm 2}$
iteration-0 slot, trained			
the same, untrained	48 / 37	80 / 73	62 / 44
ratio, trained / untrained	$5.2_{\pm 0.1}\times$ / 1.3×	$4.3_{\pm 0.2}\times$ / 1.1×	$8.0_{\pm 1.3}\times$ / 1.2×
objects with a stable iteration-0 slot (%), trained / untrained	$35_{\pm 1}$ / 3	$57_{\pm 0}$ / 15	$47_{\pm 2}$ / 24
frames in which the iteration-0 slot is the final owner, trained / untrained	$0.822_{\pm 0.002}$ / 0.020	$0.853_{\pm 0.001}$ / 0.007	$0.790_{\pm 0.011}$ / 0.127
patches in the object's majority iteration-0 cone (%), model / oracle ceiling	87.2 / 96.9	87.2 / 99.9	86.9 / 94.3
objects with a stable iteration-0 slot (%), clips exactly separable / the rest	47.7 / 32.8	62.4 / 49.9	55.9 / 44.3
seed survival after refinement, $\cos(\text{out}, \text{init})$	$0.25_{\pm 0.02}$	$0.27_{\pm 0.02}$	$0.72_{\pm 0.12}$
(2) fraction of an object's patches that a hyperplane through the origin separates in the trained projection, mean over objects			
merged / kept objects (AUROC)	0.949 / 0.984 (0.66)	0.999 / 0.999 (0.50)	0.919 / 0.971 (0.66)
objects merged (%)	41	49	56
(3) single-switch objects whose split lies in the central band $[.4, .6)$*; a uniform distribution would give 20%*			
in the central band (%)	6.2	15.3	11.0
within .1 of an end (%)	70.7	37.0	43.0
(4) owner switches per adjacent-frame transition, by object mass			
smallest / largest sixth of objects by mass	0.33 / 0.03	0.16 / 0.02	0.27 / 0.02
objects with one owner for the whole clip (%)	37.3	58.7	39.1
within-object patch pairs with one owner (%)	96.1	96.3	89.3

50% and 56% against 44%. The joint pass's cones hold most of each object but fall short of the ceiling everywhere, and exact separability of a clip raises stability without securing it. Block (4) is a further prediction of the account, kept out of the main text because it is confounded by object size; it is read below.

D.1 Identity switches become rarer as objects grow

Equation (6) makes the cost of a switch quadratic in the mass involved, so if optimization removes expensive errors first, switches should be rarer on large objects. Following each ground-truth object's owner across frames and counting switches per adjacent-frame transition in six equal-count bins of object mass, the rate falls monotonically from the smallest sixth to the largest on every dataset, from .33 to .03 on MOVi-C, .16 to .02 on MOVi-E and .27 to .02 on YTVIS (Figure 4, top; block 4). This is consistent with the m^2 weighting without isolating it, since large objects are easier for reasons unrelated to the objective; the lopsidedness prediction of block 3, the bottom row of the figure, has no such confound. The scaling does account for the gap between two readings of identity: only 37%, 59% and 39% of objects keep one owner for their whole visible life, yet counted as the objective counts, over within-object patch pairs, identity holds for 96%, 96% and 89%. The switches sit on small objects and brief excursions, where they are frequent, cheap and nearly invisible to the index.

E Qualitative results

Figure 5 adds the worst case that Figure 3 leaves out. In the two wins the object holds one slot for the whole clip in the joint attention and in the decoded mask, while SlotCurri's slot changes once or

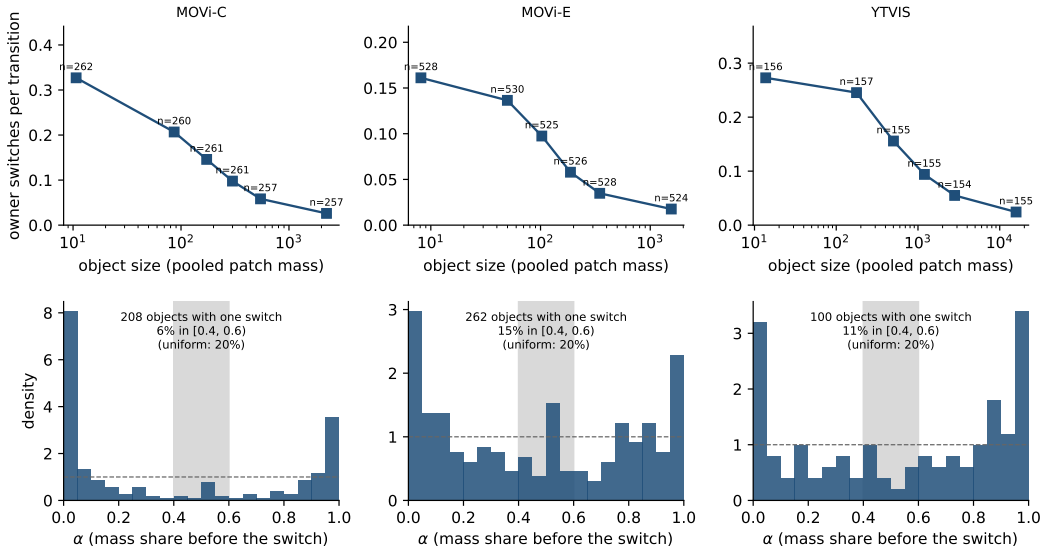


Figure 4: The two predictions of (6) on the trained models, blocks (3) and (4) of Table 5. Top: owner switches per transition by object size, six equal-count bins per dataset. Bottom: for objects with exactly one switch, the share of the object's mass before the switch; the band marks $[0.4, 0.6)$, the dashed line the uniform density.

twice. In the loss, four skiers on a snow field are missed entirely (Clip FG-ARI $-.006$ against $.851$): small objects on a uniform background are absorbed by the slots that hold the background, and no later stage can recover an object the joint pass never separated.

Figure 6 is the MOVi-C counterpart. In the win, a box that fills most of the frame is one slot for all 24 frames, while SlotCurri divides it among three slots whose colors change over the clip. In the loss, two small look-alike objects exchange slots between frames; SlotCurri holds each in its own slot. Both are the failure Proposition 2 names: objects whose patches the projection does not separate share a cell.

Figure 7 reads the MOVi-C lead per clip, and it is not broad: CoSlot is above SlotCurri on 107 of 250 clips, within $.01$ on 40 and below on 103 (sign test $p = .84$), with a median margin of zero. The $.017$ lead in Table 1 is a mean effect, larger wins than losses on a balanced count, unlike YTVIS where the lead holds on half the clips.

F THE YTVIS REFERENCE MASKS

The segmenter does most of the per-frame work on YTVIS, so how its masks are obtained and consumed is part of the method. We run SAM 3 once over the whole training split, $2{,}685$ clips and $79{,}211$ frames, prompted with YouTube-VIS's 40 category nouns, and cache the result. No ground-truth mask or track is read; prompting with the dataset's vocabulary is dataset-level knowledge. The pass is offline and paid once, and training never invokes the segmenter.

The segmenter labels objects rather than partitioning the scene, and consuming its masks needs one adjustment. A partition source such as SlotCurri assigns every patch to some channel, so an arg max over channels is defined at every patch and the largest channels are background. SAM 3 returns only the objects it finds, so a background patch is zero in every channel, and a bare arg max would assign all such patches to the first channel and fabricate one large spurious object. We therefore treat a patch as unclaimed when no channel reaches the confidence floor (Appendix G), which is every background patch, and it is dropped from the agreement term; no channel is dropped as background. We keep at most as many channels as the model has slots, the largest objects first; 98.8% of clips have no more objects than that, and matching the two counts matters because surplus columns in the agreement table are nearly empty and only dilute the index.

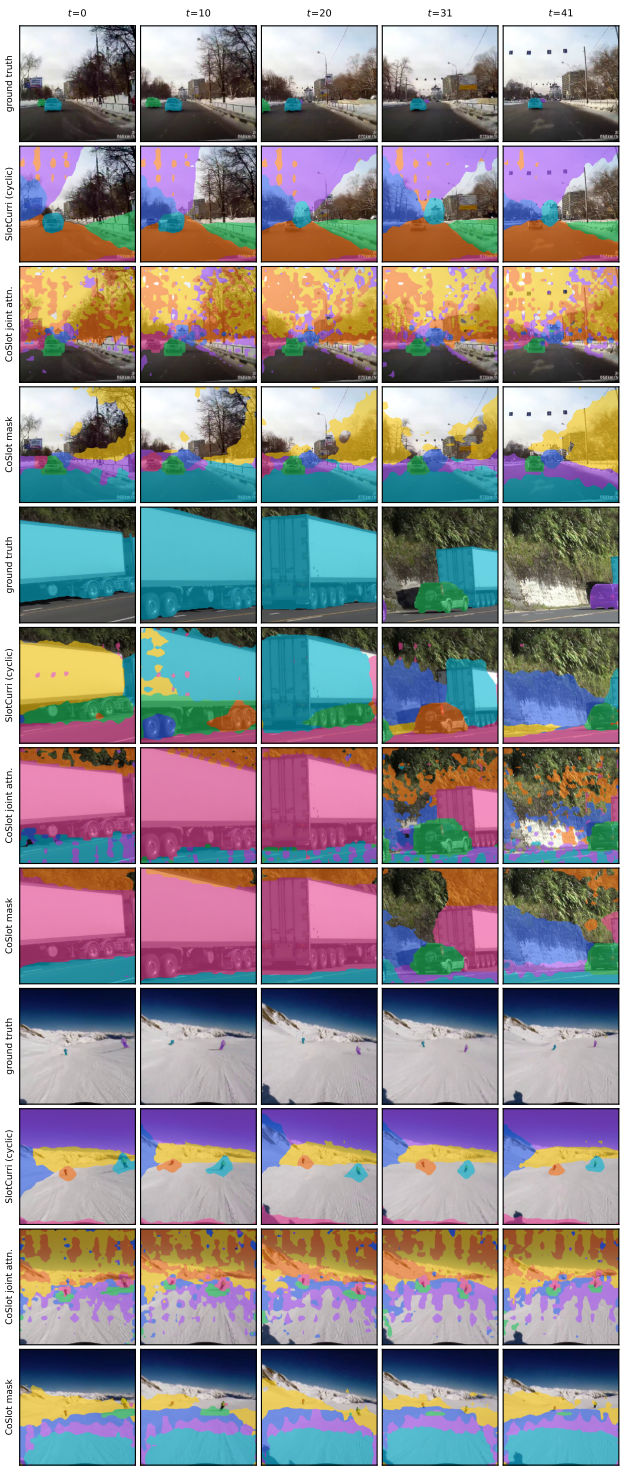


Figure 5: YTVIS: the two largest per-clip margins over SlotCurri and the largest margin against us, five frames each. Rows as in Figure 3.

Scored through our own evaluation on the 300 validation clips, SAM 3's masks reach Frame FG-ARI .853 against .488 for SlotCurri's: the segmenter's advantage is per-frame quality. Trained against SlotCurri's YTVIS masks instead, the same architecture sits at .37 Clip FG-ARI over its last three evaluations, below that source's own .44 (§4.2). Removing the agreement term leaves YTVIS at .152 against .494 (Table 6), while on MOVi-C feature reconstruction alone reaches .634: on real

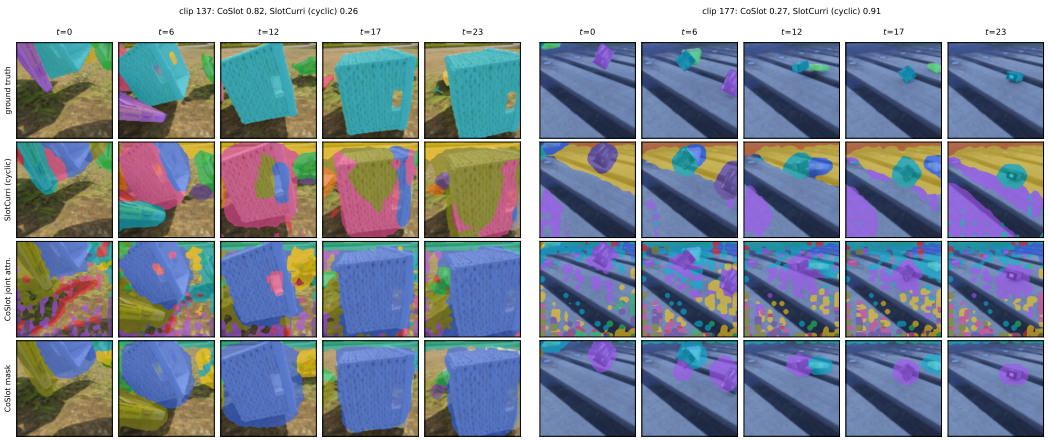


Figure 6: MOVi-C: our largest margin over SlotCurri (left) and its largest margin over us (right). Rows as in Figure 3.

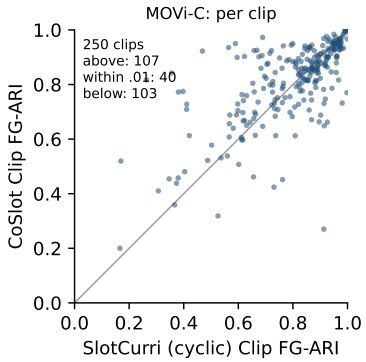


Figure 7: MOVi-C validation, one point per clip, as the left panel of Figure 2. All clips have 24 frames.

video the reference masks are where the model learns what an object is, and the agreement term is what turns them into a model.

G Training and evaluation details

Table 6: Ablation study, absolute scores. The arms of Table 2 with their scores; Δ is the change relative to the configuration with the key restored.

	MOVi-C				YTVIS			
arm	Clip FG-ARI	Δ	mBO	Δ	Clip FG-ARI	Δ	mBO	Δ
no agreement term	.6341	−.1784	.2717	−.0713	.1522	−.3421	.1127	−.2249
contiguous frames	.7574	−.0551	.2398	−.1032	.4940	+.0037	.3443	+.0118
untied per-frame attention	.8129	+.0004	.3300	−.0130	.4790	−.0152	.3314	−.0063
no slot-contrastive term	.8077	−.0048	.3342	−.0088	.5142	+.0181	.3757	+.0049
no background penalty	—	—	—	—	.5434	+.0492	.2668	−.0709

The released configuration files are the source of truth for all three settings. All three share one architecture: a frozen DINOv2 backbone (Oquab et al., 2024), ViT-S/14 at input 336 for MOVi and ViT-B/14 at input 518 for YTVIS, whose patches are projected and bound by a single slot-attention module, run first over the whole clip and then, with the same weights, over each frame's own patches, three iterations at each scope. The clip-level slots seed every frame unchanged, so no per-frame initializer and no temporal embedding exist, and a per-position MLP decoder emits the masks. Training uses Adam at 4×10^{-4} on MOVi and 1.2×10^{-3} on YTVIS after 2500 warm-up

steps, gradient clipping, and eight frames sampled from each clip. What differs between the settings is capacity and schedule: K is 11, 15 and 7 slots on MOVi-C, MOVi-E and YTVIS, the slot-count curriculum grows in proportion, and MOVi-E and YTVIS run longer and enable the background penalty. Runs are 300k steps on MOVi-C, 490k on MOVi-E and 240k on YTVIS; every reported number is from three such runs per setting. The loss weights, with reconstruction at 1, are 0.05 on the structure term in every setting, 0.5 on MOVi and 0.1 on YTVIS on the slot-contrastive term, on the agreement term 4 on MOVi and 1 on YTVIS for its copy on the joint attention and 3 everywhere for its copy on the decoded masks, and 1 on the background penalty on MOVi-E and YTVIS, held after the final slot-growth step and then annealed over 40k steps to 0.55 and 0.25; MOVi-C uses no background penalty.

Unclaimed patches, which the agreement term ignores and the background penalty charges for, arise differently on the two kinds of reference. On YTVIS the segmenter returns object tracks and nothing else, so 81% of patches are unclaimed by construction and the confidence floor, which leaves a patch unassigned when the reference's confidence in its object is below 0.8, adds 1%. On MOVi the reference is a full partition that is confident on objects and diffuse elsewhere, so the confidence floor alone leaves 83% of patches unclaimed on MOVi-C and 68% on MOVi-E, roughly the background share; no reference channel is dropped.

Three preprocessing choices decide which features a model sees without appearing in any parameter shape, and we pin all three to match the baselines: patch tokens are taken after the backbone's final normalization layer, inputs are normalized by the ImageNet statistics after the resize, and positional embeddings are interpolated to the input size. With all three fixed this way our features are identical to the baselines'.

Every metric is computed at one canonical resolution per dataset, equal to the encoder input size, which is what the baselines' own evaluation configurations use. Soft masks leave the model on the patch grid and are bilinearly upsampled to that resolution before any hard assignment is taken; reducing by $\arg\max$ on the patch grid quantizes every object boundary to the lattice and costs 0.008–0.012 Clip FG-ARI on identical predictions. Ground-truth labels are brought to the same resolution by nearest-neighbor sampling, since instance ids are categorical. This matches the baselines' own transforms, and every number we report, ours and every re-scored baseline, passes through this one function. On YTVIS, evaluation is on a central square crop of each frame, following SlotCurri and SlotContrast, whose validation transforms accept only central crops. Frames are 320 pixels on the short side and 426 to 568 wide, so the crop keeps every row and discards 25–44% of the width, and predictions and ground truth are scored inside the same window. Our YTVIS numbers are therefore comparable to the baselines re-scored here and not to full-frame YTVIS numbers reported elsewhere.

Every baseline is re-scored from its released checkpoint rather than quoted. Where the published protocol is fully specified, SlotCurri on the two MOVi datasets, our re-scoring reproduces its published mBO to within 0.003 and its Clip FG-ARI to within 0.006 on MOVi-E and 0.020 on MOVi-C, and our SRL measurement matches the figure its repository reports for that checkpoint. On YTVIS no reproduction claim is possible, because every paper scores its own split (§4); there the bar is our measurement of the released checkpoint on our split. One deviation runs against us: our measurement of SlotCurri's MOVi-C checkpoint is 0.796 against its published 0.776, so re-measuring raised the bar, and every MOVi-C margin we claim is computed against the higher value. Our SlotContrast copy is vendored in our own tree rather than run from an upstream checkout, so its fidelity to the published method is not independently attested.