VIDEO FOUNDATION MODELS BIND OBJECTS BUT DO NOT MAINTAIN THEM UNDER OCCLUSION

Anonymous authors
Paper under double-blind review

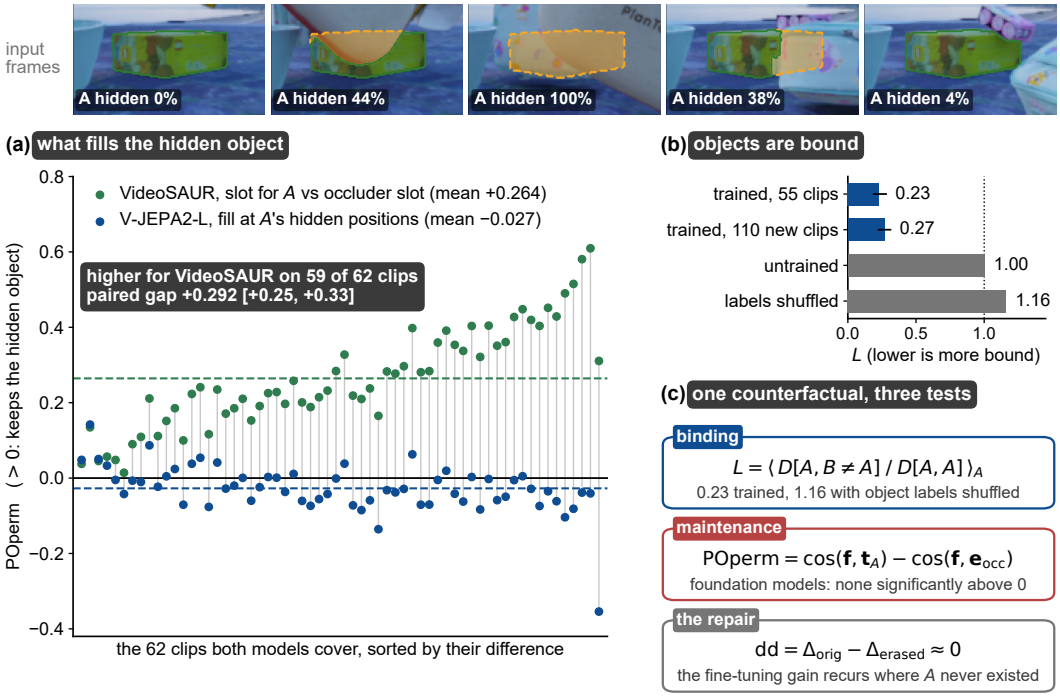


Figure 1: **Video foundation models bind objects but do not maintain them under occlusion. Top:** one occlusion event (MOVi-MC-AC); green, the visible part of object A; orange dashed, its hidden extent. **(a)** On the 62 clips both models cover, V-JEPA2-L's fill at A's hidden positions is negative on average (-0.027) while VideoSAUR's slot for A stays positive ($+0.264$), higher on 59 of 62 clips. Each model is read in its own space, and V-JEPA2-L's value here is before the calibration of §5. **(b)** The same predictor binds objects: deleting A's history damages A's future far more than other objects' (L=0.23, against 1.00 untrained and 1.16 with labels shuffled). **(c)** One context-removal counterfactual on a frozen model gives both readouts; the erased-history counterfactual shows that occlusion fine-tuning installs a history-blind prior, not tracking.

ABSTRACT

Video foundation models are widely assumed to acquire object-centric structure, yet whether that structure is causally used, and whether it survives the moment an object stops being visible, has not been tested. Decoding is blind to both: linear readouts recover object identity from frozen features, yet a probe trained to report the hidden object retrieves the occluder instead. We measure object binding inside the model's own prediction machinery: delete object A's early tokens from a frozen predictor's context, then read the per-object degradation of its future predictions. A leakage ratio L summarizes this against endpoints measured in the same pipeline, at $K+1$ forward passes per clip with no fitted parameters. In a predictive video foundation model (V-JEPA2) and a diffusion one (CogVideoX), objects are strongly bound, and in V-JEPA2 the binding is learned: $L = 0.23$–0.43 against untrained 1.00 and label-shuffled 1.16, surviving distance- and appearance-matched controls and generalizing to real video. The same models have no object perma-

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.

nence. Under occlusion the fill contrast POperm, hidden object minus occluder, is never significantly positive on our pilot scenes in five video foundation models spanning three training paradigms; where it turns positive at scale (V-JEPA2-H) it is as positive when the object never existed, and V-JEPA2's significant deficits survive subtracting the readout's own measured floor. An object-centric model with per-object recurrent state reaches $+0.264$; on the 62 clips both cover, the paired gap is $+0.292$ with the object-centric model higher on 95% of them. The obvious fix works and does not do what it appears to: two epochs of occlusion-completion fine-tuning turn permanence positive ($-0.024 \rightarrow +0.070$ over three seeds), and eight lift occluded-object identification from .369 to .522, yet a paired counterfactual in which the object never existed reproduces 96–108% of the fill gain, identifying the fix as a largely history-blind de-occlusion prior. On this class of feedforward predictors, internal binding does not imply object permanence, and the cheap fix does not install it.

1 INTRODUCTION

Video foundation models are pretrained to predict or generate video at scale, and are increasingly described as world models (Ha & Schmidhuber, 2018; Assran et al., 2025) whose representations capture intuitive physics such as object permanence (Garrido et al., 2025). The evidence for object structure is mostly decoding: probes read object positions or physical categories out of frozen latents (Zhang, 2026; Liu, 2026). Decodability establishes that the information is present in the representation. It leaves open the two questions that decide whether object structure is a property of the *model* rather than of its features: whether the model's own prediction machinery uses that structure, and whether the structure survives occlusion, when the pixels supporting it are gone.

We answer both with one instrument (Figure 1c). Object A's early-frame tokens are deleted from a frozen predictor's context; the predictor then predicts the late-frame tokens of every object, and we record how much each object's prediction degrades. The degradation matrix $D[A, B]$ gives a leakage ratio L, the damage to other objects' futures relative to A's own: $L{=}0$ when deleting A damages only A's future, $L{\rightarrow}1$ when the damage is spread over the scene. The same counterfactual, applied to an occluded region instead of an object's history, asks what the predictor fills a hidden object's position with: the object it saw before (permanence), or the occluder now in front of it. Both are interventions on the frozen predictor's context rather than on the encoder: masking is applied inside the predictor, so residual traces of A survive in the retained tokens and L measures how much the prediction machinery's use of A is confined to A, not how much A is absent from the representation.

Contributions. *An instrument.* We contribute a calibrated causal probe for object binding (§3): counterfactual context removal in prediction space, with a synthetic oracle ($L{=}0$), untrained ($L{=}1.00$) and label-shuffle ($L{=}1.16$) endpoints, at $K{+}1$ forward passes per clip, with nothing trained. *A protocol.* We also contribute the erased-history counterfactual (§6), a cheap and decisive test that any claim of learned object permanence should pass, and which overturns our own. *Findings.* Three, below, each carrying its evidence class: $^{\mathsf{R}}$ replicated on fresh held-out scenes or across ≥ 3 seeds, $^{\mathsf{S}}$ a single pass at the stated n, $^{\mathsf{E}}$ post-hoc analysis of an existing run.

Objects are bound, and the binding is learned (§4). The leakage ratio replicates at 0.267 on 110 fresh clips$^{\mathsf{R}}$, survives locality and appearance controls, and generalizes to real video and to diffusion.

Binding does not become permanence (§5). None of five video foundation models spanning predictive, diffusion and masked-autoencoding training maintains an occluded object$^{\mathsf{S}}$, against $+0.264$ for the object-centric VideoSAUR, and the gap survives on the 62 clips both arms cover. An occupant-swap counterfactual locates the mechanism: identity at an occluded position follows the *current* occupant and does not revert to the object whose history is byte-identical.

The obvious fix is a prior amplifier (§6). Two epochs of occlusion-completion fine-tuning turn held-out permanence positive ($-0.024 \rightarrow +0.070 \pm 0.012$, 3 seeds) at no binding cost this design resolves, and eight epochs lift identification .369$\rightarrow$.522$^{\mathsf{R}}$. An erased-history counterfactual then reproduces 96–108% of the two-epoch fill-similarity gain where the object never existed. The fix is real on the task, its gain history-conditional; on the permanence readout long occlusions stay broken.

2 RELATED WORK

Object-centric learning. Slot-based models build binding in by architecture and train slots to compete for image regions, from images (Locatello et al., 2020; Seitzer et al., 2023) through video (Kipf et al., 2022; Singh et al., 2022; Wu et al., 2023) and our control (Zadaianchuk et al., 2023).

Video foundation models. V-JEPA2 (Assran et al., 2025) and its predecessor V-JEPA (Bardes et al., 2024), CogVideoX (Yang et al., 2025) and VideoMAE (Tong et al., 2022) span predictive, diffusion and masked-autoencoding training. Garrido et al. (2025) report 98% zero-shot accuracy on IntPhys (Riochet et al., 2022) for V-JEPA and near chance for pixel prediction; we report no positive maintenance. A surprise readout of the same kind also fails for our baseline on *our* scenes (-0.021, §6), so the gap is not a readout-type artifact. IntPhys scenes are curated and low-clutter, ours cluttered, and POperm here falls only as severity grows, consistent with both results; the model also differs (V-JEPA, V-JEPA2), and we did not run the interpolation that would separate them.

Probing and causal analysis of frozen models. Probes find physical structure in V-JEPA2 latents (Liu, 2026); object-level masking during training forces interaction-dependent prediction (Nam et al., 2026); in small RL world models, linear probes decode object positions and steering along them changes predictions (Zhang, 2026). Outside vision, amnesic probing removes a property and measures the behavioural cost (Elazar et al., 2021), and the Othello work reads out an emergent board state and intervenes on it (Li et al., 2023). Our delta: post-hoc *causal* counterfactuals in the prediction space of frozen video models, per object, with calibrated endpoints, across paradigms.

Object permanence and intuitive physics. Benchmarks from Physion (Bear et al., 2021) to IntPhys 2 (Bordes et al., 2025) and the video-generation evaluations of Bansal et al. (2025a;b) score behaviour on curated scenarios; we score representations on ordinary occlusions and pair the negative with an architecture that passes an analogous test. Supervised trackers (Tokmakov et al., 2021) and amodal tracking benchmarks (Hsieh et al., 2023) target permanence directly; we ask whether a model trained only to predict retains objects on its own. Concurrently, Lu et al. (2026) move the camera to intervene on observability and find across 23 models that a target returning to view resumes the state it left in rather than advancing its event, a behavioural test of generated video; Liu & Chen (2026) find permanence-aligned structure, with mixed control specificity, in small transition models including *recurrent and state-space* ones, the class our scope leaves open.

3 A CALIBRATED COUNTERFACTUAL PROBE FOR OBJECT BINDING

3.1 COUNTERFACTUAL CONTEXT REMOVAL IN PREDICTION SPACE

A frozen encoder maps a clip to tokens $\mathcal{T}$, token i in time bin t_i; a frozen predictor maps a context set to predictions at target positions. Writing $c_o(i)$ for the fraction of token i's patch-tubelet covered by object o, ground-truth segmentation attributes token i to $\mathrm{obj}(i) = \arg\max_o c_o(i)$ when $\max_o c_o(i) \geq \tau = 0.5$ and sets $\mathrm{obj}(i) = \emptyset$ otherwise (Appendix D); splitting the T temporal token bins at $t^* = \lfloor T/2 \rfloor$ gives $\mathrm{early}(A) = \{i : \mathrm{obj}(i){=}A,\ t_i < t^*\}$ and $\mathrm{late}(A)$ its complement in A. The context is $\mathcal{C} = \{i : t_i < t^*\}$, background included; ablating A removes $\mathrm{early}(A)$ from it, and the targets are the late-half tokens of the K objects present in both halves. Then

$$D[A,B] \;=\; \frac{1}{|\mathrm{late}(B)|} \sum_{i \in \mathrm{late}(B)} \mathrm{dist}\big(\mathrm{pred}(\mathcal{C} \setminus \mathrm{early}(A))_i,\ \mathrm{pred}(\mathcal{C})_i\big),$$

$$L \;=\; \frac{1}{K} \sum_{A} \frac{\mathrm{mean}_{B \neq A} D[A,B]}{D[A,A]},$$

and dist is cosine distance; a relative-L_2 readout is reported alongside and agrees in ordering (per-clip Spearman 0.93, Appendix C). As a ratio, L has $L{=}1$ as a no-selectivity reference rather than an upper bound; the difference form $\mathrm{SI} = D[A,A] - \mathrm{mean}_{B \neq A} D[A,B]$ agrees in sign everywhere.

Cost. One baseline forward pass plus one per ablated object: $K+1$ encoder–predictor passes per clip, under `no_grad`, on the released checkpoint. Nothing is trained and no parameter is fitted, so L can be computed for a new model from a handful of annotated clips. Wall-clock was logged for the permanence readout: both arms of the calibration of §5 on all 56 pilot clips took 54 s (V-JEPA2-L) to 106 s (V-JEPA2-g) on one RTX 5090. POperm also needs amodal masks: synthetic corpora supply them, and on real video we obtain them by pasting occluders.

3.2 THE PERMANENCE READOUT

Occlusion events come from amodal masks. With M_A the tokens whose patch-tubelet A's amodal mask covers by more than 0.3, let $\mathrm{vis}(A) = \{i \in M_A : \mathrm{obj}(i) = A\}$ and $\mathrm{occ}(A) = \{i \in M_A : \mathrm{obj}(i) \notin \{A, \emptyset\}\}$, the positions where A is present but another object is seen, pooled over the clip; an event is an object with $|\mathrm{occ}(A)| \geq 1$ and $|\mathrm{vis}(A)| \geq 4$. We remove $\mathrm{occ}(A)$ from the full token set $\mathcal{T}$: what does the predictor fill it with? With $\bar{\mathbf{e}}(S)$ the mean encoder feature over S and $\mathbf{f}(S) = \mathrm{mean}_{i \in S}\, \mathrm{pred}(\mathcal{T} \setminus S)_i$ the fill,

$$\mathrm{POperm} \;=\; \cos\big(\mathbf{f}(\mathrm{occ}(A)),\, \mathbf{t}_A\big) \;-\; \cos\big(\mathbf{f}(\mathrm{occ}(A)),\, \bar{\mathbf{e}}(\mathrm{occ}(A))\big),$$

where the template $\mathbf{t}_A$ is $\bar{\mathbf{e}}(\mathrm{vis}(A))$, or $\bar{\mathbf{e}}$ of a random half of $\mathrm{vis}(A)$ when the other half is the control, and $\bar{\mathbf{e}}(\mathrm{occ}(A))$ is the occluder as the encoder sees it there; POperm > 0 means the hidden object wins the fill. Masking a *visible* patch of A is the sanity control: there the fill does look like A (0.80 against 0.75 at occluded positions for V-JEPA2-L, an ordering held by every trained model; Appendix D, which also shows one event in Figure 4).

Endpoints and controls. Three endpoints calibrate L: a synthetic oracle with known leakage (L=0), the same architecture untrained (L=1.00), and token-level label shuffling (L=1.16). Shuffle endpoints land slightly above 1 because L averages ratios, each with expectation $1/(1-\nu^2)$ under a gamma denominator of coefficient of variation ν (Appendix C); the bias inflates endpoints only, so the separation we claim is conservative. The untrained and shuffle endpoints come from the identical pipeline; the L=0 endpoint is analytic, since the in-model oracle, other objects painted to background, gave $L = 0.334$ on 12 clips, above the trained 0.214 (Appendix F, row 3).

Statistics and data. Statistics are clip-clustered throughout (10^4 permutations; bootstrap CIs resampling clips; Appendix E). Data: MOVi-MC-AC (Moore et al., 2025; Greff et al., 2022) with amodal masks, DAVIS-2017 (Pont-Tuset et al., 2017) and YouTube-VOS (Xu et al., 2018), contributing 165 MOVi, 48 DAVIS and 49 YouTube-VOS clips for L (Table 7) and 123 clips / 640 occlusion events at scale. All models are frozen, except the predictor fine-tuned in §6.

4 VIDEO FOUNDATION MODELS BIND OBJECTS

Strong, and learned. Deleting an object's history from the context damages that object's future prediction several times more than other objects' on average: $L = 0.229$ on the pilot set and 0.267 on 110 fresh clips, both far below their own endpoints$^{\mathsf{R}}$ (Table 1; CIs in Table 7). The untrained network of the same architecture gives 1.00 and token-level shuffling gives 1.16, with CIs disjoint from the trained value: the binding is learned. It holds under the difference-form readout, across τ, and on the large-object subset (Appendix D). The V-JEPA2 scale axis (L/H/g, Table 1) confounds capacity with other training differences and supports no conclusion.

Not locality, not appearance. A purely local predictor would produce low L without any object binding, since same-object tokens are nearby, and locality here is strong (Spearman -0.578 between a token's spatial distance to A and its degradation). It does not explain the effect: with distance partialled out over 91,219 token rows from 55 clips, same-object tokens still degrade more (partial $\rho = +0.299$, CI [+.249,+.353]), and controlling appearance similarity as well, the stronger carrier of the two, leaves $+0.161$, CI [+.127,+.197]$^{\mathsf{S}}$. About 70% of the raw same-object effect (Spearman 0.54) is locality plus appearance; the residual binding survives both. Both partials are token-row-

level with clip-clustered intervals; with the event-level partials of §6 they are where our aggregation rule is not applied to the estimand itself (Appendix E).

Real video, and a second paradigm. The effect generalizes to real video and to a diffusion video model, every setting far below its own shuffle endpoint (Table 7)[S]. DAVIS is the weak setting, and it is weak for size and motion rather than for realism, separating from YouTube-VOS at $d = 0.88$ against $d = 0.41$ for synthetic-to-YouTube-VOS and replicating at 0.624 on a disjoint split[E]. Absolute L is not comparable across paradigms, whose ablation, tokenization and timesteps differ; the claim is within-setting separation from the shuffle endpoint.

5 BINDING DOES NOT BECOME OBJECT PERMANENCE

5.1 NO POSITIVE MAINTENANCE IN ANY FOUNDATION MODEL

At occluded positions the picture inverts (Table 1, Figure 2). No tested foundation model maintains the hidden object: on the pilot scenes three of five sit significantly below zero and the other two are indistinguishable from it[S], a pattern that holds for V-JEPA2-L at scale (-0.0199, p=0.0005, 640 events / 123 clips)[R]. The one positive value is V-JEPA2-H's at scale, on one primary event per clip: $+0.043$ (p=.0003, 123 clips), but $+0.045$ (p=.0002) in the matched world where the object never existed, so it owes nothing to the object's history[S]. The two null rows are bounded: their CIs end at $+0.024$ and $+0.005$, against $+0.233$ for the object-centric control. Only VideoMAE encodes the occluder by a wide margin (similarity 0.63 to A, 0.80 to the occluder).

Three controls close the obvious escapes. First, a *trained* linear probe, the representation's best chance to reveal the hidden object, names visible objects in held-out clips (top-1 0.88–0.91 in V-JEPA2, 0.69 in CogVideoX; chance 0.21) yet at hidden positions returns the occluder more often than the hidden object in all four models, by a hair in V-JEPA2-L (0.43 against 0.42), and training makes it score *more* negative rather than less (Table 8), significantly so in two of the four models (p=.036, .013), from one probe seed. The reading is not an artifact of the fill readout. Second, an untrained network scores ≈ 0 while its probe still works, so the deficit is learned behaviour. Third, permanence degrades with occlusion severity $\mathrm{sev}(A) = |\mathrm{occ}(A)|/(|\mathrm{occ}(A)| + |\mathrm{vis}(A)|)$ in all four models: clip-wise mean Spearman runs -0.25 to -0.39, negative in 71–83% of clips[S], with POperm near zero under light occlusion and -0.098 in the heaviest quartile (Figure 2a).

The readout's own zero point. The same contrast on a *visible* patch of the same object, where the answer is unambiguously A, measures the readout's own bias (Figure 2b). With $\partial_A \subseteq \mathrm{vis}(A)$ the visible tokens with a 4-neighbour attributed to another object, so the local context is as heterogeneous as under an occluder, and $\mathbf{t}'_A = \bar{\mathbf{e}}(\mathrm{vis}(A) \setminus \partial_A)$, that contrast is $\mathrm{POperm}^{\mathrm{vis}} = \cos(\mathbf{f}(\partial_A), \mathbf{t}'_A) - \cos(\mathbf{f}(\partial_A), \bar{\mathbf{e}}(\partial_A))$, and its mean, the floor, is -0.0149 for V-JEPA2-L and -0.0306 for V-JEPA2-g (p=.0005), against -0.0005 and -0.0043 when the visible target is drawn at random: the bias grows with local heterogeneity. The corrected score $\mathrm{POperm} - \mathrm{POperm}^{\mathrm{vis}}$, both with template $\mathbf{t}'_A$, within event, leaves V-JEPA2-L and -g negative at $p \leq .002$ and V-JEPA2-H null (p=.20), as it was before the correction (Table 1)[S]. About 40% of the raw deficit in the two significant models (-0.038, -0.071 under the matched template) is the estimator; the rest is not.

5.2 AN OBJECT-CENTRIC MODEL MAINTAINS WHAT THE FOUNDATION MODELS DROP

VideoSAUR (Zadaianchuk et al., 2023) admits an analogous contrast. With $\mathbf{s}_o^{(t)}$ the slot bound to object o at occluded frame t, $\bar{\mathbf{s}}_A$ that slot averaged over the earlier frames where A is visible, and $\bar{\mathbf{s}}_{\mathrm{occ}}$ the occluder's slot averaged over the frames where it is visible, $\mathrm{POperm}_{\mathrm{slot}} = \cos\left(\mathbf{s}_A^{(t)}, \bar{\mathbf{s}}_A\right) - \cos\left(\mathbf{s}_A^{(t)}, \bar{\mathbf{s}}_{\mathrm{occ}}\right)$. It scores a *state* against the object's own earlier state, POperm a prediction against the encoder feature; both ask "object or occluder", in different spaces. It gives $+0.264$, CI [$+.233$,$+.298$], on 117 events / 62 clips, 97% of them positive[R]. It excludes 82% of the scored events, those already occluded at frame 0, where $\bar{\mathbf{s}}_A$ falls back to $\mathbf{s}_A^{(t)}$ and the first term is 1 by

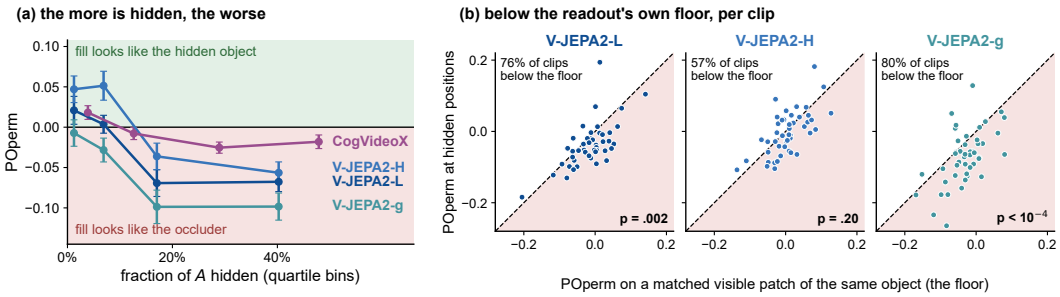


Figure 2: **The deficit grows with occlusion and, at two of three scales, sits below the readout's own floor. (a)** POperm against the fraction of A hidden, $\mathrm{sev}(A)$, in quartile bins of equal n with standard errors, for four video foundation models; bins are per object, and the clip-level statistics are in §5. **(b)** Per clip (46 clips, 158 events), POperm at A's hidden positions against the same contrast on a difficulty-matched visible patch of A, the readout's floor. Points below the diagonal are worse than the floor; p tests the within-event difference.

construction; including them gives $+0.358$ (Appendix F). Three checks rule out slot bookkeeping. The maintained content (self term 0.912) clears both the cross-slot floor (0.581) and the occluder slot (0.641), and a checkpoint trained on a different distribution reproduces the result ($+0.280$). Its own endpoint follows the label-shuffle construction of L, replacing the object's pre-occlusion identity with another object's, and gives -0.040 ± 0.001 across three slot-initialisation seeds, so the margin above it in the seeded reruns ($+0.267$) is $+0.307 \pm 0.008$, clip-$p \leq 10^{-4}$ in every seed[R]. Most audited events sit at frame 1 or 2, where $\bar{\mathrm{s}}_A$ pools one or two frames, yet the contrast holds at $+0.243$ for $t \geq 5$ (18 events), $+0.23$ to $+0.24$ above the endpoint across seeds (Appendix C)[E]. A randomly initialised VideoSAUR is no endpoint: its slots never bind (match coverage 0.512→0.119; Appendix B). And VideoSAUR's own *mask* readout is negative (-0.261 on the same events): the model that maintains the object in its representation does not reconstruct it in pixels. The floor and the second checkpoint rest on 30 and 29 audited events from separate pilot runs (Appendix B).

The arms also pair on clips: all 62 audited VideoSAUR clips are among the 109 fresh clips of the at-scale V-JEPA2 run, and within them the gap is $+0.292$, CI $[+.252,+.332]$, $p \leq 10^{-4}$, with VideoSAUR higher on 95% of clips; V-JEPA2-L sits at -0.027 $[-.044,-.012]$ on the same scenes[R]. The corpus is therefore not what separates them; the readout space still could (Limitations).

5.3 Mechanism: position-level evidence occupancy

An occupant-swap counterfactual repaints the occluder and leaves the hidden object's history byte-identical. The fill changes with the occupant alone ($1 - \cos$ of 0.038 between the two fills), its similarity to the original occluder drops (-0.014, p=.028), and the hidden object is not restored, similarity to A moving -0.015 ($p<10^{-4}$, 92 clips)[S]. Identity at an occluded position is recomputed from whatever currently occupies it, with no per-object state that this readout can detect.

6 The obvious fix is a prior amplifier

6.1 What occlusion-completion fine-tuning buys

We fine-tuned only the 22M-parameter predictor, encoder frozen, on clips disjoint from every evaluation set. With $\mathbf{f}_e = \mathbf{f}(\mathrm{occ}(A_e))$ the fill of event e and $\mathcal{E}_j$ the events of training clip j, one step per clip minimizes $\ell_j = |\mathcal{E}_j|^{-1} \sum_{e \in \mathcal{E}_j} \left[1 - \cos\left(\mathbf{f}_e, \bar{\mathbf{e}}(\mathrm{vis}(A_e))\right)\right]$ with the template fixed. This objective is one minus the *first term* of POperm on the same tokens, so a positive held-out POperm shows the objective generalizes across scenes; whether the model tracks is tested by the erased-history counterfactual below. The fine-tuning flips the metric and pays off on a task (Table 9).

Table 1: **Binding without maintenance.** Every foundation model with an L readout binds objects far below the measured endpoints (V-JEPA2-L: 1.00 untrained, 1.16 shuffled; CogVideoX: 1.12 shuffled), yet none keeps a hidden object significantly above zero on these scenes (the one positive value at scale is history-independent, §5); the object-centric model does. In parentheses: the within-event corrected score of §5, measured for V-JEPA2 with the matched template, under which the raw values are -0.038, -0.005, -0.071 (VideoSAUR, in slot space: its margin over the shuffled endpoint in the seeded reruns). Full results in Tables 7 and 8.

Model	**Training**	**Binding** $L\downarrow$	**Maintenance** POperm↑
V-JEPA2-L	predictive	**0.229**	-0.024 (-0.023)
V-JEPA2-H	predictive	0.404	$+0.007$ (-0.009)
V-JEPA2-g	predictive	0.431	-0.059 (-0.040)
CogVideoX	diffusion	0.395	-0.003
VideoMAE	masked AE	*not measured*	-0.177
VideoSAUR	object-centric	*built in (slots)*	$+\mathbf{0.264}$ ($+\mathbf{0.307}$)

Two epochs turn held-out permanence positive and eight buy the task result, identification of the occluded object as $\hat{o}_e = \arg\max_g \cos(\mathbf{f}_e, \mathbf{t}_g)$ over the scene's objects with at least four attributed tokens, $\mathbf{t}_g = \bar{\mathbf{e}}(\{i : \mathrm{obj}(i) = g\})^{\mathsf{R}}$. The gain transfers to a regime the fine-tuning never saw, strictly causal past-only prediction, and to real video in 6/6 domain×seed combinations (Figure 9a)$^{\mathsf{R}}$; the causal regime, not the identification number, is the independent evidence here.

The two-epoch point is also nearly free ($\Delta L = 0.000 \pm 0.012$ across three seeds) while the eight-epoch schedule costs $\Delta L = +0.039$ (clip-p=.35, not significant but no longer null-centred), with a single-seed sweep putting the knee at 1–2 epochs (Figure 7)$^{\mathsf{S}}$. Both are read on 16 held-out clips (baseline L CI [.151,.299]), so the design resolves a binding cost only down to about ± 0.04. All seeds share the 40 training clips, and two epochs was read off that sweep on 14 evaluation clips, which the 123 at-scale clips include beside 109 fresh ones.

6.2 The fill gain reproduces in worlds where the object never existed

The natural reading, that the model learned to track objects through occlusion, is wrong. Build an *erased-history* version of each clip in which A is painted out wherever it is visible, leaving the occluded region pixel-identical (Figure 3): the only difference between the two worlds is whether A has a history. Writing $r(w, m)$ for a readout in world $w \in \{\mathrm{O}, \mathrm{E}\}$ under model $m \in \{\text{base}, \text{fixed}\}$, the fix's gain in each world and the history sensitivity it adds are

$$\Delta_w \;=\; r(w, \text{fixed}) - r(w, \text{base}), \qquad \mathrm{dd} \;=\; \Delta_{\mathrm{O}} - \Delta_{\mathrm{E}}.$$

The two accounts predict opposite things: tracking needs the object to have existed ($\Delta_{\mathrm{E}} \approx 0$, $\mathrm{dd} \approx \Delta_{\mathrm{O}}$), a history-blind prior does not ($\Delta_{\mathrm{E}} \approx \Delta_{\mathrm{O}}$, $\mathrm{dd} \approx 0$).

The second account wins (Table 2; Figure 5): $\Delta_{\mathrm{E}} \approx \Delta_{\mathrm{O}}$ at every scale, and dd is zero for L and H (90% CI [$-.008$,$+.003$] for L); only the largest shows a detectable $+0.008$, about 7% of its gain$^{\mathsf{R}}$, which an event-level estimator puts at $+0.005$. What the fix injects is generic object content: against a position-matched reference the shift toward A is $+0.062$ and toward an *uninvolved* object $+0.047$, not a significant difference.

The baseline arm of the same counterfactual corrects our negative result in the other direction: four readouts (the fill and three encoder-side ones, Appendix C) show an original-minus-erased difference of $+0.025$ to $+0.048$ ($p<10^{-4}$; d=0.52 on the fill readout, 123 paired events). The hidden object does leave a trace, subdominant and losing the contrast readout that pits A against the occluder, and it is specific to the erased object rather than a side effect of repainting (difference-of-differences against an uninvolved object $+0.011$ and $+0.030$, $p<10^{-4}$)$^{\mathsf{S}}$. Two traces must be told apart: the encoder-side trace is *larger* untrained ($+.068$ to $+.096$) and is therefore attention-level pixel leakage$^{\mathsf{R}}$, while whether the predictor-side trace is learned stays open, since an untrained predictor's fill has cosine 0.02 to every reference and its null difference (p=.83) cannot tell.

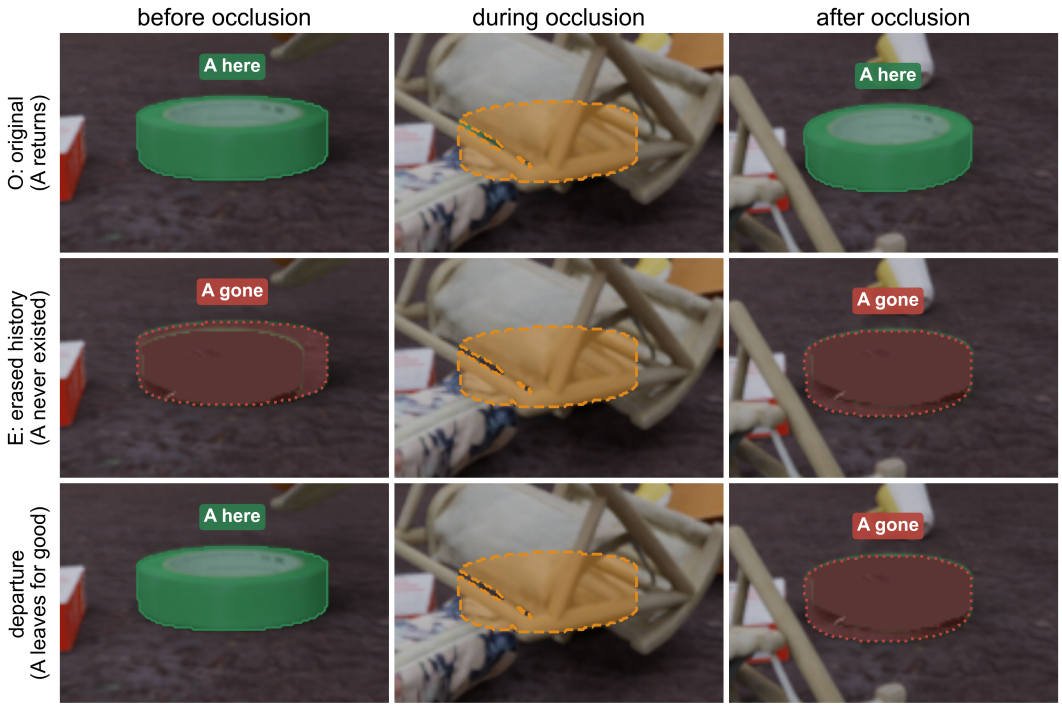


Figure 3: The paired counterfactual worlds, built from one clip. **O**: the original. **E**: A is painted out wherever it is visible, so it never existed. **Departure**: A is visible before the occlusion and never returns. The hidden region is pixel-identical in all three during the occlusion, so any difference in the fill there can only come from A's history. Green: A visible; orange dashed: its amodal position while hidden; red dotted: where A has been erased.

Table 2: **The fix does not need the object's history.** Fill similarity to A before and after occlusion-completion fine-tuning, in the original world (O) and the world where A never existed (E); 123 paired events per scale. The gain is as large in E as in O.

	V-JEPA2-L		**V-JEPA2-H**		**V-JEPA2-g**	
Fill similarity to A	O	E	O	E	O	E
baseline	0.776	0.737	0.751	0.732	0.757	0.729
fine-tuned	0.876	0.846	0.858	0.837	0.875	0.843
gain	+0.101	+0.109	+0.106	+0.106	+0.119	+0.114
gain in E / gain in O	1.08		0.99		0.96	

What the fix is, precisely. Four results keep the fix from being metric gaming. It adds no false reports. In erased-history worlds, false reports of A do not rise from the baseline's .31 (-0.009, p=.72, 122 clips; the event-level rate moves $+0.005$; Appendix E). The task gain is conditional on the object having existed: the original-minus-erased gap in reporting A grows .058→.156, a factor of 2.7. In a departure world the fine-tuned model reports A *less* often than the baseline (.451→.361), so its gap to the original world grows .041→.131, a factor of 3.2. And on a violation-of-expectation readout (Appendix C) the baseline fails (-0.021: it does not expect the object back) and the fix passes ($+0.012$, p=.0005, 68% of clips)[S].

The account that fits all of this is narrow and mechanical: fine-tuning installs a history-blind de-occlusion prior, "put object-like content behind occluders", which suppresses occluder dominance and lets the baseline's weak echo win the argmax, so the task gain depends on history while the fill gain does not. Its one direct prediction, that events with a larger baseline echo flip to A more often, holds in direction only (Spearman $+0.18$ over the 61 events the baseline got wrong, p=.176)[E]. The gain is real and adds no tracking.

6.3 THE FAILURE REGIME: LONG OCCLUSIONS, AT ANY DEPTH

The fix has a boundary, and it is temporal. Severity mixes two per-event axes. With h_t and v_t the numbers of tokens of $\mathrm{occ}(A)$ and $\mathrm{vis}(A)$ in time bin t and $\mathcal{B}_h = \{t : h_t > 0\}$ among the bins $\mathcal{B}$ where A is present, $\mathrm{depth} = |\mathcal{B}_h|^{-1} \sum_{t \in \mathcal{B}_h} h_t/(h_t + v_t)$ is how much of A is hidden while hidden and $\mathrm{duration} = |\mathcal{B}_h|/|\mathcal{B}|$ for how much of its presence: fractions, not pixel areas or frame counts, correlating weakly (Spearman 0.23). Within an event the baseline deficit is full-size at onset and does not deepen (Appendix C), so duration effects below are differences between events, not decay within one. The partial Spearman correlations below control the other axis (event-level, cluster-bootstrap CIs, 640 events / 123 clips; Table 5 has the rest). The baseline deficit grows with both (depth -0.20 [$-.28$,$-.13$], duration -0.13 [$-.21$,$-.04$])[E]. The gain depends on duration alone (-0.324 [$-.393$,$-.251$]; -0.345 on 56 events / 14 clips), not on depth ($+0.045$, CI containing zero at both sizes)[R] (Figure 8). After the fix the depth term is where it was (-0.22 [$-.30$,$-.14$])[E] and the duration term has grown to -0.577 [$-.638$,$-.512$]: the fix lifts every depth by about the same amount, and long occlusions least. On 56 events / 14 clips the gain falls $+0.115/+0.103/+0.033$ by duration tertile, against the expectation we had recorded, and the ordering survives a severity-independent template control[S].

Training-signal allocation does not explain the long-occlusion deficit: reweighting the loss $10\times$ toward long occlusions leaves the long-occlusion gain unmoved across three seeds[R]. A history-blind prior predicts exactly this: it is anchored on what was visible a moment ago, so there is no track to extend, and duration is the axis on which a prior cannot substitute for state. The boundary belongs to the permanence readout: on identification the gain runs the other way by duration tertile ($+0.03/+0.14/+0.15$) while the fill's margin shift is flat ($+0.017/+0.017/+0.019$), so which argmax flips depends on where the baseline margin sat[E].

7 DISCUSSION

Object structure inside the video foundation models we test is real, causal and learned, and it does not amount to object permanence: on the same 62 clips V-JEPA2-L's contrast is negative (-0.027) and an object-centric model's is $+0.26$. The corpus is ruled out as the cause; the readout space remains open (Limitations). The practical deliverable is a control: occlusion-completion fine-tuning improves every permanence number we have and is still not tracking, which the erased-history counterfactual shows by reusing the evaluation clips with one object painted out; any claim of learned object permanence should pass it. The next measurement is the same contrast on a video model that carries state.

Limitations. Every foundation model tested is a feedforward predictor without persistent per-object state, and the positive control is one object-centric architecture, so that side of the dissociation is $n{=}1$ in architectures and rests on 117 audited events. Real video is two datasets with pasted occluders, leaving natural occlusion statistics untested, and absolute values are readout-dependent, so differences and not levels carry the claims: "no permanence" means no positive maintenance with a subdominant trace, not amnesia. POperm's own bias is measured (§5); the correction is a difference of two contrasts on the same event, so it inherits both their variances, and it is single-pass at each scale. POperm compares cosines of mean features, and the number of tokens averaged rises with severity, so part of Figure 2a's gradient may belong to the estimator; per-token cosines would separate the two and were not computed. The two arms of the dissociation are matched on clips and on endpoint construction but still read different spaces, and no readout we have scores a slot model and a feedforward predictor identically.

AI USE STATEMENT

Generative AI tools were used for the following tasks on the policy's required-disclosure list: proposing and refining hypotheses; designing and giving feedback on methodology and experiments, including several of the controls reported here; implementing methods, that is the analysis

and plotting code; cleaning and reformatting data, in the form of re-analyses of stored per-event records; developing the conceptual framing of the results; and interpreting results. On the same list, we did not use generative AI to generate synthetic data sets: the video corpora are public releases, and the counterfactual worlds were produced by deterministic scripts. Formulating or proving mathematical claims, assisting with translation, and qualitative or thematic data analysis do not apply to this work. Beyond that list, generative AI was also used to search for related work, to draft sections of the paper and to polish its writing.

We reviewed the AI-assisted work as follows. Every number in the paper is traced to a stored artifact (Appendix B) and was re-checked against that artifact by script, which is how four transcription errors were caught; reading the implementations caught a contaminated positive control (Appendix F, row 16). Controls proposed with AI assistance were run and their outcomes recorded whether or not they favoured our claims; the corrections that followed are listed in Appendix F. Every citation added with AI assistance was checked against the publisher or arXiv record before inclusion. The authors set the research direction, decided which experiments to run and when to stop, and made every positioning, retraction and scoping decision. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work measures publicly released models on public video corpora. It involves no human subjects, collects no personal data, and releases no new imagery. The synthetic corpus is MOVi-MC-AC (Moore et al., 2025), rendered with Kubric (Greff et al., 2022) and released under CC BY 4.0; DAVIS-2017 (Pont-Tuset et al., 2017) and YouTube-VOS (Xu et al., 2018) are used under their published terms, and we release only derived per-event measurements and the scripts that produce them, never re-hosted frames. The six models are public checkpoints used forward-only, with one exception: a 22M-parameter predictor fine-tuned on occlusion completion, which changes how an existing model fills occluded regions and confers no capability its base model did not have.

Two risks are specific to a negative result of this kind. It can be over-read as a claim about video foundation models in general, so the scope is stated wherever the claim is made: every model we test is a feedforward predictor without persistent per-object state, and the object-centric contrast rests on one architecture. It can also be over-read as a claim that occlusion handling in deployed systems is unsafe, which we do not test; nothing here is an evaluation of a deployed tracker. Working the other way, the fix we report improves an occlusion benchmark while adding no object tracking, and we say so explicitly rather than reporting the benchmark gain alone. The record of what we retracted along the way (Appendix F), including a positive control we found to be contaminated and corrected downward, is part of the submission. All released material is anonymised for review.

REPRODUCIBILITY STATEMENT

Everything needed to repeat the measurements is specified in the paper or shipped with it. The probe is defined in Section 3: token attribution and the context-removal counterfactual, then the permanence readout with its sanity control, then the calibrated endpoints, the matched controls and the statistics protocol. Appendix C gives the distance readouts and the derivation of the endpoint bias, Appendix D the attribution audit and the τ sweep, and Appendix E the statistical protocol: clip-clustered permutation tests, cluster bootstraps, the de-pseudo-replication rule, the multiplicity treatment, and the rule for reporting multi-seed families. Appendix B maps every claim in the paper to the script that produced it and to the stored artifact that holds its numbers.

The anonymous supplementary material contains those artifacts, one JSON file per run, together with every script that produced them, the plotting scripts that redraw each data figure from them, the re-analysis scripts, among them the audit that recomputes the positive control on events with pre-occlusion history, and the download and preprocessing scripts for the three corpora. Run from the package on a CPU, every re-analysis reproduces its stored artifact exactly and every data figure is pixel-identical to the one printed here. The three figures that show video frames need the corpus, since no frames are shipped, and table entries are read from the artifacts Appendix B names. Models

are public checkpoints used forward-only, so the only training to reproduce is the 22M-parameter predictor fine-tune: 40 training clips, at most eight epochs, three seeds, with the seeds and the schedule reported alongside each number that depends on them. Compute is modest by construction, since the probe costs $K+1$ forward passes per clip and fits no parameters.

Two limits on what reproduction can establish. This work was not preregistered: "replication" here means re-running a fixed analysis on held-out scenes, results carried over from an earlier pass are marked $^{\mathsf{S}}$, and post-hoc analyses are marked $^{\mathsf{E}}$. And the corrections we made are part of the record rather than hidden by it, so a reader who reruns the positive control will obtain the audited $+0.264$ of Table 8 up to slot-initialisation seed ($+0.267 \pm 0.009$ over three seeds), not the $+0.358$ that the unaudited event set gives. VideoSAUR draws its initial slots at random and the original run did not seed them (Appendix F); the endpoint script takes an explicit `--torch_seed`.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations (ICLR)*, 2025a. arXiv:2406.03520.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025b.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. arXiv:2404.08471.

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Florian Bordes, Quentin Garrido, Justine T. Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. IntPhys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, et al. Kubric: A scalable dataset generator. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. arXiv:2203.03570.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Cheng-Yen Hsieh, Kaihua Chen, Achal Dave, Tarasha Khurana, and Deva Ramanan. TAO-amodal: A benchmark for tracking any object amodally. *arXiv preprint arXiv:2312.12433*, 2023.

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.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations (ICLR)*, 2023.

Hung Ming Liu. Probing the latent world: Emergent discrete symbols and physical structure in latent representations. *arXiv preprint arXiv:2603.20327*, 2026.

Yang Liu and Yuming Chen. Event-conditioned diagnostics of kinematic, contact, and object-permanence structure in passive object-state world models. *arXiv preprint arXiv:2606.28455*, 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. arXiv:2006.15055.

Jinpeng Lu, Dexu Zhu, Haoyuan Shi, Linghan Cai, Guo Tang, Yinda Chen, Jie Cao, Duyu Tang, Yi Zhang, Yong Dai, and Xiaozhu Ju. Current world models lack a persistent state core. *arXiv preprint arXiv:2606.20545*, 2026.

Alexander Moore, Amar Saini, Kylie Cancilla, Doug Poland, and Carmen Carrano. Training for X-ray vision: Amodal segmentation, amodal content completion, and view-invariant object representation from multi-camera video. *arXiv preprint arXiv:2507.00339*, 2025.

Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-JEPA: Learning world models through object-level latent masking. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.11389.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. IntPhys 2019: A benchmark for visual intuitive physics understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5016–5025, 2022. doi: 10.1109/TPAMI.2021.3083839.

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.

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.

Pavel Tokmakov, Jie Li, Wolfram Burgard, and Adrien Gaidon. Learning to track with object permanence. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2203.12602.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. SlotFormer: Unsupervised visual dynamics simulation with object-centric models. In *International Conference on Learning Representations (ICLR)*, 2023.

Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. YouTube-VOS: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2408.06072.

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. arXiv:2306.04829 (VideoSAUR).

Xinyu Zhang. What do world models learn in RL? probing latent representations in learned environment simulators. In *ICLR Workshop on World Models: Understanding, Modelling and Scaling*, 2026. arXiv:2603.21546.

A NOTATION

Table 3: Symbols, in order of first use. Bold lower case is a feature vector; roman names are estimators.

Symbol	Meaning	Where
Tokens and objects		
$\mathcal{T}$, i, t_i	the clip's tokens; a token; its time bin	§3
T, t^*	number of time bins; the split $\lfloor T/2 \rfloor$	§3
o, A, B, g	objects; A the ablated or hidden one, g a gallery candidate	§3, §6
$c_o(i)$, τ	fraction of token i's patch-tubelet covered by o; attribution threshold (0.5)	§3
$\mathrm{obj}(i)$, $\emptyset$	the object token i is attributed to; unattributed	§3
$\mathrm{early}(A)$, $\mathrm{late}(A)$	A's tokens with $t_i < t^*$ and with $t_i \geq t^*$	§3
$\mathrm{pred}(S)_i$	the frozen predictor's output at token i given context S	§3
$\mathbf{e}_i$, $\bar{\mathbf{e}}(S)$	encoder feature of token i; its mean over the set S	§3
Binding		
$\mathcal{C}$	context: every token with $t_i < t^*$, background included	§3
K	number of objects present in both halves	§3
dist	cosine distance $1 - \cos$ (relative L_2 in Appendix C)	§3
$D[A, B]$	degradation of B's late predictions when $\mathrm{early}(A)$ leaves the context	§3
L, SI	leakage ratio; its difference form	§3
Permanence		
M_A	tokens whose patch-tubelet A's amodal mask covers by more than 0.3	§3
$\mathrm{vis}(A)$, $\mathrm{occ}(A)$	tokens of M_A where A, or another object, is seen	§3
$\mathbf{f}(S)$	fill: mean prediction over S when S is removed from $\mathcal{T}$	§3
$\mathbf{t}_A$	template of A: $\bar{\mathbf{e}}(\mathrm{vis}(A))$, or of a random half	§3
POperm	fill contrast, hidden object minus occluder	§3
$\mathrm{sev}(A)$	occlusion severity, the hidden share of $\mathrm{vis}(A) \cup \mathrm{occ}(A)$	§5
∂_A, $\mathbf{t}'_A$	visible tokens of A bordering another object; template without them	§5
$\mathrm{POperm}^{\mathrm{vis}}$	the same contrast on ∂_A, the readout's floor	§5
$\mathbf{s}_o^{(t)}$, $\bar{\mathbf{s}}_A$, $\bar{\mathbf{s}}_{\mathrm{occ}}$	VideoSAUR slot of o at frame t; pooled slot of A, of the occluder	§5
$\mathrm{POperm}_{\mathrm{slot}}$	the contrast in slot space	§5
The fix		
e, A_e, $\mathbf{f}_e$	an occlusion event; its hidden object; its fill $\mathbf{f}(\mathrm{occ}(A_e))$	§6
$\mathcal{E}_j$, ℓ_j	events of training clip j; its fine-tuning loss	§6
$\mathbf{t}_g$, $\hat{o}_e$	gallery template; the object identified at event e	§6

Symbol	Meaning	Where
w, m, $r(w, m)$	world (O original, E erased history); model (base, fixed); a readout	§6
Δ_w, dd	gain of the fix in world w; history sensitivity $\Delta_{\mathrm{O}} - \Delta_{\mathrm{E}}$	§6
h_t, v_t	tokens of $\mathrm{occ}(A)$, of $\mathrm{vis}(A)$ in time bin t	§6
$\mathcal{B}$, $\mathcal{B}_h$	bins where A is present; where it is hidden	§6
depth, duration	hidden share while hidden; hidden share of the bins	§6
Appendix readouts and statistics		
W, θ, d_e, p_W	probe map, temperature, encoder width, probe probability	App. C
R_A, P, σ_w	A's re-emergence tokens; the context before them; surprise in world w	App. C
X_A, Y_A, μ, a, ν	numerator, denominator of L's ratio; their shuffle mean; gamma shape; coefficient of variation	App. C
j, J, x_j, $\bar{x}$	clip index; number of clips; within-clip mean; mean over clips	App. E
N, b, $\varepsilon_j^{(b)}$	permutation draws; draw index; random signs	App. E
p, d, ρ	p-value; Cohen's d; Spearman correlation	throughout
R, S, E	evidence classes, written as superscripts: replicated, single pass, post-hoc	§1

B EXPERIMENT INVENTORY

Every number in the paper is backed by a stored JSON artifact in the released repository; the table gives claim → script → artifact. The rows below cover every claim in the main text, including the checks added after the first submission draft (P3–P6, P8).

Table 4: Claim, producing script and stored artifact for every result in the paper.

Claim (section)	Script	Artifact
L, synthetic and real, both readouts (§4)	`phaseF5`	`phaseF5_predmetric_*` (3 scales × 2 readouts × 3 datasets)
Endpoint calibration, synthetic oracle	`phaseF1`, `phaseF6`	`phaseF5_f1_oracle_*`, `phaseF6_calibration_*`
Attribution audit, τ sweep, big-object subset	`phaseF7`	`phaseF7_audit`, `phaseF5_*_tau03/tau07/top2/top3`
L replication on fresh scenes	`phaseF30`	`phaseF30_L_at_scale`
Diffusion arm, timestep sweep	`phaseF8`	`phaseF8_cogvideox_t*` (3 timesteps, shuffle, DAVIS)
Distance-matched control	`phaseF37`	`phaseF37_distance_control`
Appearance-matched control	`phaseF41`	`phaseF41_appearance_control`
POperm, V-JEPA2 three scales, untrained	`phaseF11`	`phaseF11_amodal_perm*` (3 scales, untrained)
POperm, diffusion	`phaseF12`	`phaseF12_cogvideox_perm_t100`
Trained linear probe	`phaseF13(b)`	`phaseF13_perm_probe_*`, `phaseF13b_cogvideox_probe_*`
POperm, masked autoencoder	`phaseF20`	`phaseF20_videomae_permanence{,_ext}`
Positive control, cross-slot floor, second checkpoint	`videosaur`	`videosaur_permanence*` (pilot, ext, movi-e, x-slot, untrained)
Domain reversal on real video	`videosaur`, `phaseF25`	`videosaur_permanence_davis/ytvos`, `phaseF25_real_permanence*`
Occlusion-completion fine-tuning	`phaseF15`	`phaseF15_finetune_permanence`
Downstream identification	`phaseF16`, `phaseF33`	`phaseF16_downstream_utility`, `phaseF33_downstream_at_scale`
Binding cost, trade-off curve	`phaseF17`, `phaseF18`	`phaseF17_modularity_preserv.`, `phaseF18_tradeoff_curve`
Causal past-only regime	`phaseF21`, `phaseF32`	`phaseF21_causal_permanence`, `phaseF32_causal_at_scale`

Claim (section)	Script	Artifact
Multi-seed stability	`phaseF27`, `phaseF28`	`phaseF27_multiseed`, `phaseF28_downstream_multiseed`
Mechanism at scale: depth, duration, gating	`phaseF29`, `phaseF23`, `F23b`, `F35`	`phaseF29_expanded_mech.`, `phaseF23_duration_vs_degree`, `F23b_interaction`, `F35_cluster_sens.`
Within-event decay, visibility control	`phaseF34`	`phaseF34_decay_curve`
Duration reweighting, 3 seeds	`phaseF24`	`phaseF24_heavy_weighted_fix*` (3 seeds)
Depth/duration, baseline, gain and post-fix	`P8` (on F23, F29 rows)	`P8_depth_duration_decomposition`
Erased-history counterfactual, 3 scales	`phaseF36`	`phaseF36_object_memory*`, `F36_supplement_dd*` (3 scales)
Template specificity, position control	`phaseF38`	`phaseF38_template_spec.`, `F38_supplement_location`
Hallucination rate, gain decomposition	`phaseF39`, `F42`	`phaseF39_hallucination_rate_*`, `F42_gain_decomposition`
Departure world, violation of expectation	`phaseF40`	`phaseF40_departure_voe`
Occupant swap	`phaseF43`	`phaseF43_occupant_swap`
Frame-0 audit of the positive control (§5)	`P3`	`P3_videosaur_frame0_audit`
Clip-paired dissociation, hallucination test (§5, §6)	`P4` (on F29, videosaur rows)	`P4_paired_dissociation`
Readout floor on matched visible patches (§5)	`P5`	`P5_readout_calibration_*` (3 scales × random/boundary)
Identity-shuffle endpoint for VideoSAUR (§5)	`P6`	`P6_videosaur_endpoint_*` (3 seeds, untrained arm)
Positive control by event frame (§5)	`P10` (on `videosaur`, `P6` rows)	`P10_videosaur_lag_strata`
Recognition-gating control (null)	`phaseF26`	`phaseF26_recognition_gate`
Statistics hardening	`P2`	`P2_stats_hardening`, `stats_cosine/relative`

C METRIC DETAILS

Depth, duration and time course at scale. At scale (n=640) the baseline deficit splits into an occlusion-*depth* term and a *duration* term (Table 5). They are additive and independent: no interaction, no mediation through fragment recognisability (partial $= +0.040$), no gating (high-recognisability -0.206 against low -0.143)[S]; an earlier two-channel account built on n=53 dissolved at 12× the sample (Appendix F). Within an event the deficit does not deepen frame by frame. It is at full size at onset and then drifts slightly *toward* recovery ($+0.0017$ per frame bin, 90% CI [$+.0011$,$+.0024$]), while a fully visible control object over the same frames stays flat (slope -3×10^{-5}, CI spanning zero). This is occlusion-specific, not a forgetting curve[E] (Figure 9b).

Distance readouts. With u the prediction under ablation and v the prediction under the full context at the same target token, the cosine readout is $1 - \cos(u, v)$ and the relative readout is $\|u - v\|/\|v\|$; both are scale-free, which is what makes L comparable across model widths. The two give different absolute values and the same ordering: on the pilot set the trained model is $L = 0.229$ [.187,.283] under cosine and 0.428 [.394,.462] under relative, with a per-clip rank agreement of Spearman 0.93; the unnormalized raw readout gives 0.426 [.394,.459]. The untrained network is at 1.001, 1.000 and 1.000 under the three readouts, so the endpoint does not depend on the choice.

Linear probe. For a feature $\mathbf{x}$ and the clip's object templates $\mathbf{t}_g = \bar{\mathbf{e}}(\{i : \mathrm{obj}(i) = g\})$ over its attributed objects g,

$$p_W(g \mid \mathbf{x}) \;=\; \mathrm{softmax}_g\big(\cos(W\mathbf{x}, \mathbf{t}_g)/\theta\big),$$

with $W \in \mathbb{R}^{d_e \times d_e}$ (d_e the encoder width) initialised at the identity and the temperature θ learned from 0.1. It is trained by cross-entropy to name the object of visible tokens (at most 300 per clip) in one half of the clips and tested on the other half, two folds, so every clip is tested once. On held-out visible tokens its top-1 is the visible accuracy of §5; at an event it scores $p_W(A \mid \mathbf{f}) - p_W(\mathrm{occ} \mid \mathbf{f})$ with $\mathbf{f} = \mathbf{f}(\mathrm{occ}(A))$ and occ the object most often attributed within $\mathrm{occ}(A)$.

Violation-of-expectation readout. In the original world O and the departure world, where A never returns, let R_A be A's visible tokens after its last occluded time bin in O and $P = \{i : t_i < \min_{i' \in R_A} t_{i'}\}$ the strictly earlier context. With $\sigma_w = 1 - \mathrm{mean}_{i \in R_A} \cos\big(\mathrm{pred}_w(P)_i, \mathbf{e}_i^w\big)$ the surprise in world w, the readout is $\sigma_{\mathrm{dep}} - \sigma_{\mathrm{O}}$, positive when the model expected A back; it is scored on the primary event of each clip with $|R_A| \geq 2$ and $|P| \geq 8$.

Trace readouts. The baseline trace of §6 is read four ways, each as the original-minus-erased difference of a cosine to $\mathbf{t}_A$: the predictor's fill $\mathbf{f}(\mathrm{occ}(A))$ ($+0.039$), the encoder feature at $\mathrm{occ}(A)$ ($+0.026$), and the pooled ($+0.025$) and best-matching ($+0.048$) encoder tokens outside A's extent in the occluded time bins, which are pixel-identical across the two worlds.

Positive control by event frame. Of the 117 audited VideoSAUR events, 44 sit at frame 1, 33 at frame 2, 22 at frames 3–4 and 18 at frame 5 or later. By stratum, $\mathrm{POperm}_{\mathrm{slot}}$ is $+0.251$ [$+.201$,$+.304$], $+0.285$ [$+.233$,$+.334$], $+0.254$ [$+.174$,$+.333$] and $+0.243$ [$+.166$,$+.308$] (17 clips in the last), each $p < 10^{-4}$. The margin over the identity-shuffle endpoint, per seed, is $+0.30$ to $+0.34$ in the first three strata and $+0.242$, $+0.230$, $+0.234$ at frame 5 or later ($p \leq .0002$). The self term $\cos(\mathbf{s}_A^{(t)}, \bar{\mathbf{s}}_A)$ is 0.88–0.93 in every stratum. Script `P10_videosaur_lag_strata`, stored rows only.

Difference form. $\mathrm{SI} = D[A,A] - \mathrm{mean}_{B \neq A} D[A,B]$ is reported alongside L in every artifact. SI has no ratio bias and no denominator instability, and it agrees in sign with L in every setting; L is the headline because it is dimensionless and its endpoints are interpretable.

Why the endpoints exceed 1. L averages per-object ratios X_A/Y_A, with $X_A = \mathrm{mean}_{B \neq A} D[A,B]$ and $Y_A = D[A,A]$. Under label shuffling no object is special, so $\mathbb{E}X_A = \mathbb{E}Y_A = \mu$, but the denominator is small and noisy; for Y_A independent of X_A, Jensen's inequality gives $\mathbb{E}[X_A/Y_A] = \mu\,\mathbb{E}[1/Y_A] > 1$. If further Y_A is gamma-distributed with shape a and scale μ/a, so that its coefficient of variation is $\nu = a^{-1/2}$, then $\mathbb{E}[1/Y_A] = a/\big((a-1)\mu\big)$ and

$$\mathbb{E}\left[\frac{X_A}{Y_A}\right] \;=\; \frac{a}{a-1} \;=\; \frac{1}{1-\nu^2}, \qquad a > 1.$$

A coefficient of variation of 0.2–0.45 gives 1.04–1.25; inverted, the measured shuffle endpoints (1.03–1.16, Table 7) correspond to $\nu = 0.17$–0.37. The per-object matrices from which ν could be measured directly were not stored, so this is a consistency check rather than a fit. The untrained endpoint has a stabler denominator and lands at 1.001. The bias moves endpoints up, never the trained value down, so every "far below the endpoint" statement in the paper is conservative.

D ATTRIBUTION-QUALITY AUDIT

Token attribution keeps a token only if one object covers at least τ of its patch-tubelet. On MOVi this assigns 11.0% of tokens on average (range 2.4–27.9%; 901 of 8192 tokens per clip belong to some object, of which the strictly pure patches are 6.2%), with a mean of 45 tokens (median 30) for the smallest object in a clip. Low coverage costs statistical power and cannot manufacture object selectivity, because unassigned tokens enter neither the ablation set nor the targets.

L is stable across the threshold: $\tau = 0.5/0.7/0.9$ gives $0.229/0.262/0.283$ (n=$55/54/51$ clips, overlapping CIs), so the mild increase tracks the shrinking token budget. Small objects are not driving the result: restricting to the four largest objects per clip gives 0.228 [.172,.304] against 0.229 for all objects. For the permanence readout the visible-patch control is the corresponding check: masking a visible patch of the target object yields a fill similarity of 0.80 to that object



Figure 4: One real occlusion event (MOVi-MC-AC, cropped to the event region). Green outline: the visible part of object A. Orange dashed: A's amodal ground-truth extent while it is hidden, the tokens we remove from the context and ask the predictor to fill. Percentages give the fraction of A's amodal mask that is covered. Events of this shape supply the 640 occlusion events used at scale.

against 0.75 at occluded positions for V-JEPA2-L, an ordering held by every trained model tested, so the readout can express "fills the object" when the evidence is there.

E STATISTICS PROTOCOL

Domain reversal. Reversing the domains tests the remaining alternative, that VideoSAUR wins because MOVi is its training distribution: real video with pasted occluders is in-domain for the foundation model and out-of-domain for VideoSAUR. The signs hold, with VideoSAUR at $+0.106$ on YouTube-VOS and $+0.094$ on DAVIS (11 and 9 events, mask readout still negative) while the foundation model still forgets. This is directional support and not a main anchor. Out-of-domain binding fails on about 80% of clips, so VideoSAUR's real-video column is conditional on binding success while the foundation model's is not, and both DAVIS arms fail the correction of Appendix E at q=0.05. On the foundation-model side DAVIS is a positive instance and YouTube-VOS a null ($+0.005$, p=.65), and a direct two-sample test does not separate them (diff $= -0.022$, p=0.12, CI [$-.047$,$+.004$]), so we claim no domain dependence[S].

Clips are the unit of analysis. Every p value in the paper is a clip-clustered permutation test with at least 5×10^3 draws (10^4 except one duration analysis whose p we do not quote): objects and events are aggregated within a clip first, and the sign or label permutation is applied at the clip level. Confidence intervals are cluster bootstraps over clips (2000 resamples). Correlations reported across objects are de-pseudo-replicated by aggregating within clip before ranking. The partial correlations of §4 and §6 are the exception: they are computed on token rows or events with cluster-bootstrap intervals, and Table 5 gives clip-aggregated values beside the event-level ones. Written out, with x_j the within-clip mean of a per-event quantity over J clips and $\bar{x} = \frac{1}{J}\sum_j x_j$, the one-sample test is a two-sided sign flip,

$$p \;=\; \frac{1}{N}\sum_{b=1}^{N} \mathbf{1}\left[\left|\frac{1}{J}\sum_{j=1}^{J} \varepsilon_j^{(b)} x_j\right| \geq |\bar{x}|\right], \qquad \varepsilon_j^{(b)} \overset{\text{iid}}{\sim} \text{Unif}\{-1,+1\},$$

and a count of zero is written $p < 10^{-4}$ or $p \leq 10^{-4}$ for $N = 10^4$ (the P-series scripts use $(\text{count}+1)/(N+1)$ instead); correlation and two-sample tests permute labels instead of signs. Intervals are percentile intervals of the statistic recomputed on clips drawn with replacement. A partial Spearman correlation regresses the ranks of each variable on the rank of the controlled one by least squares and correlates the residuals. An early version of this project reported object-level p values; the correction moved that headline from p=0.0002 to p=0.012 on the run it was measured on, and is entry 5 of Appendix F. The equivalent clip-clustered value for the run reported here is p=0.010 (Table 8).

Protocol variants of the false-report test. Three variants were run: the primary event of each clip with candidates from the whole clip (122 events), the same with cover-template candidates (122), and every event with whole-clip candidates (639 events / 122 clips). The paper reports the last. After fine-tuning, the rate of naming the erased object falls .434 → .352 and .434 → .328 in the two primary-event variants and moves .311 → .316 event-level (-0.009 clip-level) in the reported one, so the choice does not favour the claim.

Depth and duration. Table 5 gives every partial Spearman behind §6 at both sample sizes. These intervals use 5000 clip resamples, not 2000. Clip aggregation moves the duration terms of the gain by at most 0.03; it moves the other terms further, and at 56 events it flips the sign of the depth term of the gain, whose interval contains zero either way. At 56 events neither baseline term excludes zero, so the baseline decomposition in §6 rests on the 640-event set alone and carries class E.

Table 5: Partial Spearman of V-JEPA2-L's permanence readout on each occlusion axis, controlling the other, for the baseline, the fine-tuning gain and the model after fine-tuning. Depth: mean fraction of A covered in the time bins where it is occluded. Duration: fraction of A's time bins in which it is occluded at all. Event-level estimate with cluster-bootstrap 95% CI, then the clip-aggregated estimate. Artifacts: `P8_depth_duration_decomposition`, which reproduces `F35_cluster_sensitivity` for the gain and the post-fix rows.

		640 events / 123 clips		56 events / 14 clips	
state	axis	event [CI95]	clip	event [CI95]	clip
baseline	duration	-0.128 [$-.214$,$-.039$]	-0.207	-0.194 [$-.443$,$+.097$]	-0.115
	depth	-0.203 [$-.281$,$-.126$]	-0.240	-0.126 [$-.439$,$+.216$]	-0.077
gain	duration	-0.324 [$-.393$,$-.251$]	-0.345	-0.345 [$-.535$,$-.108$]	-0.374
	depth	$+0.045$ [$-.028$,$+.122$]	$+0.024$	-0.014 [$-.279$,$+.230$]	$+0.245$
after fix	duration	-0.577 [$-.638$,$-.512$]	-0.710	-0.713 [$-.811$,$-.519$]	-0.810
	depth	-0.219 [$-.298$,$-.136$]	-0.346	-0.224 [$-.498$,$+.089$]	-0.124
Spearman(duration, depth)		$+0.225$		$+0.315$	

Multiplicity. The paper reports 42 tests. They are not exchangeable, spanning different readouts, corpora and model families, so we do not apply one global correction; instead every claim carries its evidence class, and headline claims additionally carry a scale replication or a multi-seed check. We do report one correction in full. Over the 16 tests that carry a headline or a control claim, Benjamini–Hochberg (Benjamini & Hochberg, 1995) at q=0.05 has critical p=.036 and 12 of the 16 survive; at q=0.10 the critical value is .070 and 15 do. No claim in this paper rests on any of the four that fail, and we flag them in place rather than dropping them: DAVIS real baseline (p=.046), the DAVIS vector arm (.052), the CogVideoX probe (.070) and the amplifier event-test (.176; §6). Each is reported in the text as directional, and each supports a claim that a replicated result already carries. Restricted to the eight permanence tests of Table 8, the correction changes no verdict: the five $p<10^{-4}$ rows stay at adjusted 1.6×10^{-4}, V-JEPA2-L's p=.010 becomes .013, and the two null rows stay null (adjusted .469).

Multi-seed families report per-seed values and the spread; p values are not pooled across seeds that share an evaluation set, because those tests are not independent. Where a per-seed test is underpowered (the pilot downstream evaluation at 14 clips gives $p = .043/.084/.063$) the paper cites the at-scale test instead and uses the seeds only for stability.

F HONEST CORRECTIONS

Every claim withdrawn during this project, what killed it, and what replaced it. Two of these were found by controls we designed against our own headline results (rows 9 and 11), two by re-auditing old scripts after an unrelated confound turned up (rows 10 and 12), one by reading the positive control's implementation (row 16), one by an audit that traced every reported number back to its producing script (row 17), one by re-running a stored result under a fixed seed (row 18), three by reading the producing code and artifacts line by line while preparing follow-up experiments (rows 19–21), one by writing every estimator out against its code (row 22), three by re-reading the earlier reviews against the text (rows 23–25), and one by a final consistency pass (row 26).

Table 6: Every claim withdrawn during the project, what killed it, and what replaced it.

#	Withdrawn claim	What killed it	Replacement
1	Binding measured as representation displacement	Displacement never touches the predictor	Prediction-space counterfactual
2	Trained/untrained ratio $= 786478\times$	Near-zero denominator	Calibrated endpoints, L 0.21 vs 1.00
3	Bijective-relabel shuffle; graying oracle endpoint	Bijection preserves the partition; graying is out of distribution, and on 12 clips the graying oracle gave $L = 0.334$ [.259,.418], above the trained 0.214, failing the separation gate it was built to pass	Token-level shuffle; analytic synthetic oracle
4	Permanence in absolute-cosine / reappearance space	Centroid bias; scene completion drowns it	Amodal contrast against the occluder
5	Object-level p-values	Pseudo-replication (E5: p .0002 $\rightarrow$ .012)	Clip-clustered permutation
6	A general synthetic-real gap	Cross-model test: a vit-L special case	Dataset difficulty; DAVIS is the hard case
7	"Modularity buffers occlusion" ($r = -0.20$)	8 seeds: 12% significant; sign flips on CogVideoX	Weak trend $r \approx -0.13$, not claimed
8	Gap explained by object density	Top-K big-object control: L 0.17–0.21	Size and motion, not density
9	Two-channel mechanism (mediation, gating)	$12\times$ expansion (n=640) dissolves both	Additive depth + duration terms
10	Event-level correlations as the duration evidence	Pseudo-replication	Cluster bootstrap over clips, clip-aggregated values alongside (Table 5)
11	**"The fix teaches tracking"**	Erased-history counterfactual: 96–108% of the gain in worlds without the object	History-blind prior amplifier
11b	"The baseline retains zero trace"	Same counterfactual, baseline arm: $+.025$ to $+.048$	Subdominant trace, loses the contrast
12	At-scale n accounting and pairing	Filename collision: 49/640 pairs matched across clips	Re-run with parent-qualified keys; all directions survive
13	"Forgetting is domain-dependent"	Difference-of-significance fallacy; direct test p=0.12	DAVIS positive instance, YouTube-VOS underpowered
14	"The fix injects A-specific content"	Position confound: the erased baseline already shows $+0.136$	Generic object content (A $+0.062$ vs. uninvolved $+0.047$, n.s.)
15	"The scene-level trace is learned memory"	Untrained encoder trace is *larger* ($+.068$ to $+.096$)	Attention pixel leakage; whether the predictor-side trace is learned is open (row 23)
16	Positive control $= +0.358$, 99.5% of objects, content margin $+0.403$	Frame-0 events (82% of the set) compare the slot vector with itself: the pre-occlusion identity falls back to the current frame, so self-similarity is 1 by construction	Audited subset, t>0 only: $+0.264$ [$+.233$,$+.298$], 117 events, 97% positive, content margin over the cross-slot floor $+0.317$; the dissociation holds with a smaller effect
17	False-report rate $+0.005$, p=.88	No test producing that p exists: the script stores rates only, and the rate was event-level while our stated rule aggregates within clip first	Clip-clustered test on the same stored rows: -0.009, CI [$-.055$,$+.038$], p=.72, 122 clips. The claim (no hallucination) is unchanged; the sign of the point estimate is not

#	Withdrawn claim	What killed it	Replacement
18	Implicit in every VideoSAUR number: that a run is reproducible	Slots are drawn from `torch.randn` and the script seeded only NumPy, so the same (clip, object, frame) event moves by up to 0.376 between runs and the admitted count ranges 106–111 across three seeded reruns where we had reported 117	Clip-aggregated values are stable: $+0.2674 \pm 0.0085$ over three slot seeds, containing the reported $+0.264$. Per-event values are not reproducible without a seed and are no longer presented as if they were
19	The slot readout written as single frames, $\cos(\mathbf{s}_A^{(t)}, \mathbf{s}_A^{(t_0)}) - \cos(\mathbf{s}_A^{(t)}, \mathbf{s}_{\mathrm{occ}}^{(t)})$	Reading the code while building a follow-up: both references are pooled, A's slot over its visible frames before t and the occluder's over the frames where it is visible	The pooled form of §5; every reported number was always the code's
20	Duration partials labelled -0.345 at n=640 and -0.324 at n=56, and "clip-aggregated estimates within 0.03"	The stored cluster-sensitivity artifact has the two sample sizes the other way round, and at 56 events clip aggregation moves the depth term of the gain from -0.014 to $+0.245$	Labels corrected; Table 5 now gives event-level, CI and clip-aggregated values at both sizes
21	Occupant swap reported as $\Delta\text{fill} = 0.038$, $p<10^{-4}$	The quantity is $1 - \cos$ between two fills, never negative, and the test is a symmetric sign-flip permutation, which such a quantity passes by construction	The change is reported without a p; the signed evidence is the drop in similarity to the original occluder (-0.014, p=.028) and toward A (-0.015)
22	Four definitions: the context written $\bigcup_o \mathrm{early}(o)$, the fine-tuning fill written with that context, the occlusion split described per frame, and the endpoint bias written $k/(k-1)$ with k the number of objects	Writing each estimator out against its code: the context is every early-half position, background included; the fill removes only $\mathrm{occ}(A)$ from the full token set; the split is pooled over the clip; and k in the bias is a gamma shape, not an object count, with the coefficient-of-variation range a simulation input rather than a measurement	The formulas of §3 and §6 as the code computes them, and the bias derived as $1/(1-\nu^2)$ and used as a consistency check (Appendix C). Every reported number was always the code's
23	The predictor-side trace is learned: untrained difference -0.000, p=.83	The untrained predictor's fill has cosine 0.02 to every reference, so no difference could appear; the control cannot test learnedness	Whether the predictor-side trace is learned is stated as open; the encoder-side trace remains pixel leakage
24	Scoring all 56 pilot clips takes under a minute per model	The logged runs took 54 s, 81 s and 106 s for V-JEPA2-L, -H and -g, and they time the permanence readout, not L	The logged times and what they cover
25	Every endpoint of L comes from the identical pipeline	The L=0 endpoint is analytic; the in-pipeline oracle failed its gate (row 3)	Untrained and shuffle endpoints in-pipeline, the oracle stated as analytic
26	POperm is never significantly positive in five video foundation models	On one primary event per clip at scale, V-JEPA2-H's baseline contrast is $+0.043$ (p=.0003, 123 clips, `F36_supplement_dd_vith`)	Scoped to the pilot scenes; the at-scale exception is reported in §5 with its erased-world value, $+0.045$, which shows it does not depend on the object's history

Table 7: Leakage ratio L (lower is tighter binding). Endpoints are measured in the same setting where shown, not assumed, and the oracle is analytic (§3); Figure 6 plots the same values. Evidence classes as defined in §1.

Data	Model	L (cos)	CI95	L (rel)	endpoint (shuffle / untr.)	class
MOVi (55 clips)	V-JEPA2-L	**0.229**	[.19,.28]	0.428	1.16 / 1.00	R
MOVi (110 fresh)	V-JEPA2-L	0.267	[.226,.312]	–	–	R
MOVi	V-JEPA2-H	0.404	–	0.547	–	S
MOVi	V-JEPA2-g	0.431	–	0.560	–	S
MOVi	CogVideoX	0.395	–	–	1.12 (own)	S
YouTube-VOS (49)	V-JEPA2-L	**0.323**	[.25,.40]	0.536	1.05 (own)	S
DAVIS (48)	V-JEPA2-L	**0.612**	[.51,.72]	0.728	1.14 / 1.00	R
DAVIS	V-JEPA2-H	0.518	–	0.686	–	S
DAVIS	V-JEPA2-g	0.573	–	0.717	–	S
DAVIS	CogVideoX	0.594	–	–	1.03 (own)	S
synthetic oracle	–	0.000	–	0.000	perfect-binding endpoint	S

Table 8: Permanence under occlusion, in full. All rows share the corpus (MOVi-MC-AC) and the sign convention but not the space they are read in: prediction space after context removal for V-JEPA2 and CogVideoX, slot space for VideoSAUR, encoder-side for VideoMAE, which has no forward prediction to ablate. Levels are therefore not comparable across blocks; the within-block signs and the paired test of §5 carry the claim. Event sets differ with each model's coverage and binding filters; the artifacts mapped in Appendix B give the counts. CI95 is a clip-clustered bootstrap; p is a clip-clustered permutation test (10^4 draws); probe columns are a trained linear readout, the best chance to reveal the hidden object.

Model	Paradigm	State	POperm	CI95	clip-p	probe (occ./A)	class
V-JEPA2-L	predictive	none	-0.024	$[-.042,-.006]$	.010	0.43 / 0.42	R
V-JEPA2-H	predictive	none	$+0.007$	$[-.009,+.024]$	.414	0.48 / 0.35	S
V-JEPA2-g	predictive	none	-0.059	$[-.080,-.038]$	$< 10^{-4}$	0.48 / 0.38	S
CogVideoX	diffusion	none	-0.003	$[-.010,+.005]$	.469	0.34 / 0.22	S
VideoMAE	masked AE	none	-0.177	$[-.201,-.153]$	$< 10^{-4}$	–	R
VideoSAUR movi-c	obj.-centric	per-obj.	$\mathbf{+0.264}$	$[+.233,+.298]$	$< 10^{-4}$	–	R
VideoSAUR movi-e	obj.-centric	per-obj.	$+0.280$	$[+.226,+.332]$	$< 10^{-4}$	–	S
VideoSAUR, mask	obj.-centric	per-obj.	-0.261	$[-.335,-.185]$	$< 10^{-4}$	–	R
VideoSAUR, shuffled	endpoint	–	-0.040	$\pm.001$ (3 seeds)	–	–	R
untrained V-JEPA2-L	–	none	$+0.002$	$[-.001,+.005]$	n.s.	–	S

G ADDITIONAL FIGURES AND TABLES

The three below are plotted from the stored artifacts by `F_paper_appendix_figures.py`.

Table 9: Effects of occlusion-completion fine-tuning (predictor only, 22M parameters, encoder frozen). All rows are held-out clips; classes as in §1. Other eight-epoch rows and the occluder-answer rate are in the released artifacts (Appendix B).

Readout	Setting	before → after	test	class
POperm	2 epochs, 3 seeds	$-0.024 \rightarrow +0.070$	clip-$p \leq .0002$ each	R
POperm	640 ev./123 clips	$\Delta = +0.072$	clip-$p < 10^{-4}$	R
POperm	causal, 122 clips	$+0.008 \rightarrow +0.031$	clip-$p < 10^{-4}$	R
POperm, Δ only	real video, 2×3 seeds	$\Delta + 0.025$ to $+0.061$	6/6, $p \leq .0083$	R
Identification acc.	8 epochs, 640 ev.	$.369 \rightarrow .522$	clip-$p < 10^{-4}$	R
Leakage L	2 epochs, 3 seeds	$\Delta L = 0.000 \pm 0.012$	n.s.	R
Fill gain, erased world	3 model scales	96–108% of orig.	diff-in-diff. ≈ 0	R
False report, erased world	639 ev., 122 clips	-0.009	p=.72	S
False report, departure	1 run, 122 clips	$.451 \rightarrow .361$	gap to O $\times 3.2$	S
Long-occl. gain	dur. tertile 3	$+0.033$; rew. $+0.025$	$10\times$ rew.: flat, 3 seeds	R

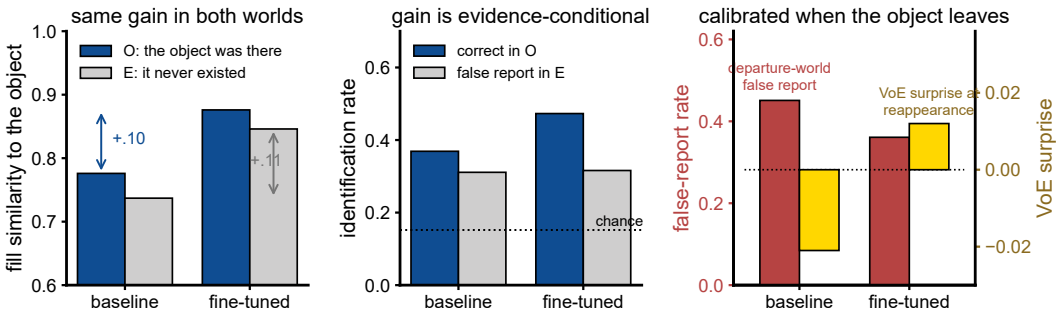


Figure 5: The fix is a prior amplifier. **Left**: fine-tuning adds the same fill similarity in the original world (O) and the erased-history world (E), so it adds no history sensitivity; the O–E offset already in the baseline is the subdominant echo. **Middle**: the gain is evidence-conditional, correct identifications rising in O while false reports in E stay flat (chance dashed), widening the O–E gap .058 to .156. **Right**: in the departure world the fine-tuned model reports the object *less* often, and the violation-of-expectation readout turns positive. Middle and right use the same runs as Table 9.

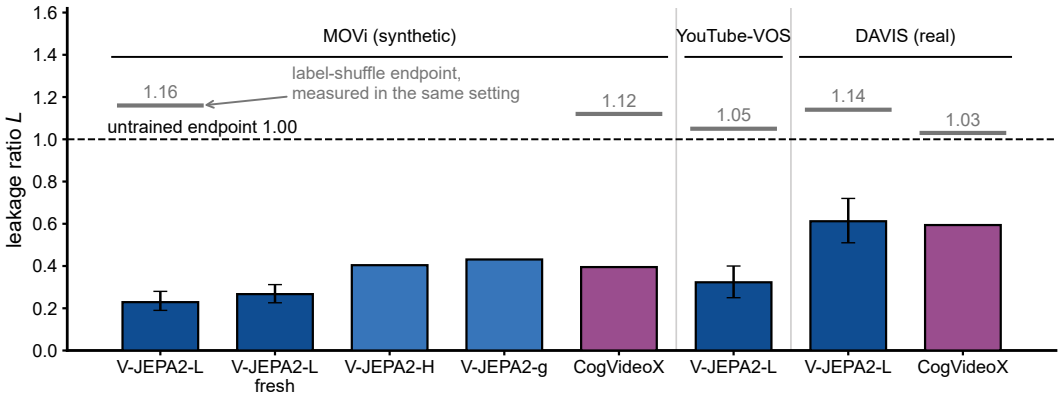


Figure 6: Object binding in prediction space. Leakage ratio L (cosine readout, lower is tighter binding) with 95% clip-bootstrap CIs where available, against calibrated endpoints: synthetic oracle L=0, untrained L=1.00, and the label-shuffle endpoint measured in each setting (grey). Every trained setting sits far below its own shuffle endpoint: in synthetic video, in two real-video datasets, and in a diffusion video model. Absolute L is not comparable across paradigms (§4); the within-setting separation is the claim.

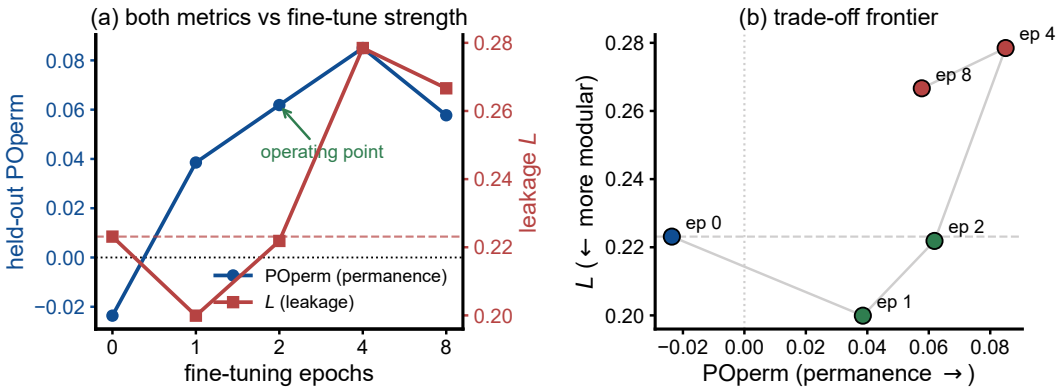


Figure 7: Inducing permanence costs binding only when the fine-tuning is pushed. **(a)** Held-out POperm (left axis) and leakage L (right axis) against fine-tuning epochs; the dashed line is the baseline L. **(b)** The same five points as a frontier. One to two epochs buy positive permanence at L at or below baseline; four epochs raise L to 0.278, and by eight epochs permanence has fallen back to $+0.058$, so the extra leakage buys no further permanence. Single seed; the multi-seed check at the operating point is $\Delta L = 0.000 \pm 0.012$ (§6).

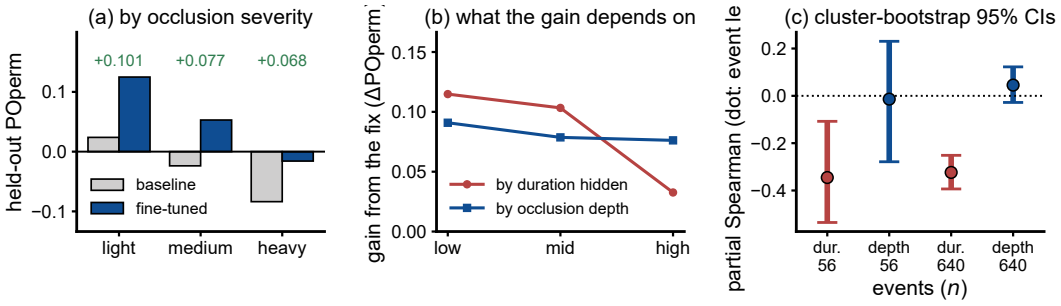


Figure 8: What the fix improves, and what it does not. **(a)** By occlusion severity, 56 events: the gain is positive in all three bins (overall clip-$p = 0.0004$) and largest where the occlusion is lightest; heavy occlusions end at roughly zero. **(b)** Splitting severity into duration and depth (56 events, tertiles), the gain falls steeply across duration tertiles and is flat across depth tertiles. **(c)** Partial Spearman correlations between the gain and each axis, with cluster-bootstrap 95% CIs at both sample sizes (dots are the event-level estimates): duration excludes zero in both datasets, depth contains it in both.

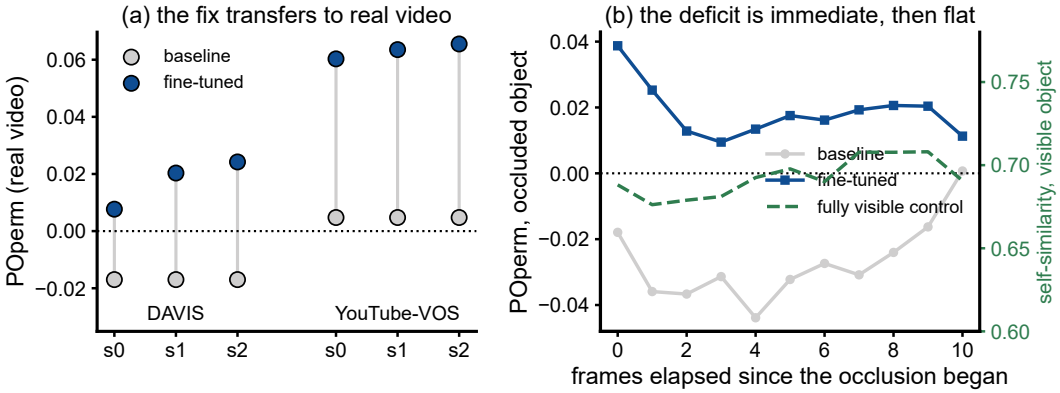


Figure 9: **(a)** The fix transfers to real video with pasted controlled occluders: 2 domains $\times$ 3 seeds, every run positive and significant, with the DAVIS baseline negative and the YouTube-VOS baseline at zero. **(b)** Within an occlusion event the baseline deficit is at full size at onset and then drifts slightly toward recovery (Appendix C), not a forgetting curve, while the fine-tuned model stays positive throughout; a fully visible control object measured over the same frames (right axis) shows no decay, which makes the effect occlusion-specific. Bins with fewer than 150 events are dropped.