POSITION: INTERACTION UNDERSTANDING REQUIRES MORE THAN OBJECT-CENTRIC REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Interaction understanding should be an explicit target of representation learning for agents that act in the physical world. A system can maintain object identities and recognize an interaction category while leaving open how a particular relationship is developing: whether a grasp is being established or released, whether a handover has begun, when the current contact will end. We argue that the state of the interaction between particular participants, and how it changes, should be an explicit learning target rather than an assumed byproduct of object perception or interaction recognition, and that progress should be judged by whether that state transfers to new questions about the event. A pilot study with SAM4D, a small readout trained on frozen SAM 3 features and told which two participants to attend to, illustrates the distinction. Telling the readout both participants raises relation Recall@3 by 19.39 percentage points over an actor-only query from the same features. On 15 events from eight videos, however, the recognition-trained representation shows no reliable advantage over controls with the same head, non-visual inputs, and training budget at predicting how much time is left in the current event, even though the same classifier, replayed on the same frames, still ranks the recorded label in its top three on 44.80% of them, averaged over videos. Two human annotators who did not see the automatic labels both record 9 of the 15 events, with start and end times that differ from the automatic ones by a median of 0.2 s; on these nine events the label-only duration prior still has the lowest error. In this pilot, recognizing an interaction and knowing its current state come apart. Whether objectives that target this state transfer better is untested; we specify the matched comparisons of learning objectives that would decide it, and what each outcome would mean for the position.

1 INTRODUCTION

A person is holding a cup. A perception system may locate both participants, maintain their identities, and recognize *holding(person, cup)*. An assisting robot may also need to know whether the cup was just picked up or has been held steadily, whether a second hand has made contact and a handover has begun, whether the grip is tightening or letting go, and when the current contact will end. These are questions about this particular occurrence, an event in the sense of event perception (Zacks & Swallow, 2007; Kurby & Zacks, 2008), which we call *event-level* questions; answering them is what we mean by interaction understanding, and the participants and the category can remain unchanged while the answers change. Figure 1 shows this in an existing video: within one *holding* event the supplied participants and the recorded label stay fixed while the time left in the event changes.

The practical question for representation learning is which part of training should learn the information that event-level questions need. One option trains the representation with a general objective, such as predicting missing or future parts of a video or recognizing interaction labels, and relies on it to keep enough information. Each downstream task then learns on its own to extract what it needs; current world models take this route (Assran et al., 2025; Agarwal et al., 2025; Zhou et al., 2025). The other option trains the representation to tell apart the interaction states that several tasks need, such as a grip tightening or letting go, even when the objects and the interaction label stay the same. Its costs are labels for these states, a fixed list of states that may be too narrow, and a

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.

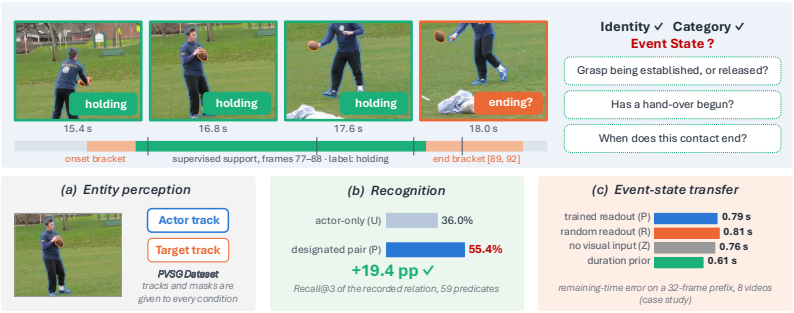

Figure 1: **Case study.** Top: one *holding* event; the supplied contours (actor blue, target orange) and the recorded label stay the same while the interaction develops, and brackets mark the automatic endpoint labels. Identity and category leave the questions on the right open. Bottom: (a) what the pilot supplies, (b) what specifying the participants improves, (c) what the pilot leaves open.

shared state that a controller built for one task may not need. The kind of model does not settle this choice: object-centric video models such as SAVi and SlotFormer, neural relational inference, and predictive world models already represent dynamics and relations beyond identity and location (Kipf et al., 2022; Wu et al., 2023; Kipf et al., 2018; Assran et al., 2025; Zhou et al., 2025). What matters is what the training objective asks a representation to keep, that is, its learning target.

In this paper, an *object-centric* representation is one whose learning target is object identity, location, and persistence over time, the information a tracker needs. The *more* in our title is the state of the interaction between particular participants and how it changes.

Our position. Interaction understanding should be an explicit learning target, not an assumed byproduct of object perception or interaction recognition. A representation should preserve the state of the interaction between particular participants and how it changes, and progress should be judged by whether that state transfers to new event-level questions. Three terms carry the statement. *Participant-grounded* means tied to particular tracked entities rather than to categories, as in the binding problem (Greff et al., 2020; Locatello et al., 2020). *Interaction state* means the distinctions needed to answer the event-level questions at hand (Section 2.1). *Transfer* means answering a question the representation was not trained on, within a stated limit on adaptation data and compute (Kornblith et al., 2019; Zamir et al., 2018). The position has two parts. The first says what to aim for and how to check it: a representation should expose the state of a particular event, not only its category, and a model trained with a general objective meets this part if it does so (Section 2). The second is a hypothesis about how to get there: that objectives which explicitly target this state transfer better than general objectives.

We design SAM4D, which is a small readout(Alain & Bengio, 2017; Belinkov, 2022) trained on frozen SAM 3 features (Carion et al., 2026). The pilot tests the last of the questions above: when the current contact will end. First, telling the readout both participants matters: it raises Recall@3 on the recognition task, the fraction of recorded relation labels ranked in the top three, by 19.39 percentage points over an actor-only query (Section 3.1). Second, a small head that sees only past frames, trained on the frozen recognition-trained representation to predict how long the current event will continue, shows no reliable advantage over controls with the same head and training budget on 15 events from eight videos. The recognition classifier, replayed on the head's own inputs, still ranks the recorded label in its top three on 44.80% of them, averaged over videos, and on the nine events that two human annotators both record, the label-only duration prior still has the lowest error. In this pilot, recognizing an interaction and knowing its current state come apart.

Overall, this paper makes three moves: it states the target grounding in particular participants, sensitivity to relational change, and temporal coherence, with transfer to new event-level questions as the test (Section 2); it illustrates the distinction with the pilot, where specifying the participants improves recognition but the recognition-trained representation shows no reliable advantage on a remaining-time question about the same events (Section 3); and it sets out the alternative views (Section 4) and the matched comparisons whose outcomes would support, qualify, or overturn the position (Section 5).

2 WHAT SHOULD A REPRESENTATION PRESERVE?

2.1 OBJECTS PROVIDE GROUNDING; EVENTS SPECIFY THE TARGET

An *interaction category*, such as *holding*, groups occurrences. An *event* is a particular occurrence involving particular entities over an interval; the same participants can take part in several events, so their identities alone do not specify the occurrence (Zacks & Swallow, 2007; Kurby & Zacks, 2008). An *interaction state* denotes the distinctions needed to answer a stated family of questions about that occurrence, given the available observations; it can be continuous, spread across several tokens, or only partly observable, and uncertainty and missing evidence are part of the target. Binding research motivates participant correspondence, and event perception motivates maintained event references.

Existing objectives already cover parts of the target. Object-centric and relational-dynamics models such as SAVi, SlotFormer, and neural relational inference learn object representations and interacting dynamics from video, and relational structure from object trajectories (Kipf et al., 2022; Wu et al., 2023; Kipf et al., 2018). Interaction and event recognition—human–object interaction, video relation, scene-graph, anticipation, state-change, and progress tasks—supervises roles, temporal organization, and unfolding events (Chao et al., 2018; Gupta & Malik, 2015; Kim et al., 2021; Shang et al., 2017; 2019; Ji et al., 2020; Cong et al., 2021; Yang et al., 2023; Damen et al., 2022; Grauman et al., 2022; Furnari & Farinella, 2019; Girdhar & Grauman, 2021; Becattini et al., 2020). Scene graphs, affordances, vision–language–action models, and predictive world models connect representations to planning and control (Hughes et al., 2022; Gu et al., 2024; Bahl et al., 2023; Brohan et al., 2023; Kim et al., 2025; Black et al., 2025; Assran et al., 2025; Zhou et al., 2025; Mur-Labadia et al., 2026). Each is scored on its own task, and transfer studies find that an upstream score can misjudge the benefit for a new use (Kornblith et al., 2019; Ericsson et al., 2021; Majumdar et al., 2023; Hu et al., 2023).

2.2 OBJECT-TASK SUCCESS LEAVES THE INTERACTION STATE OPEN

A representation can succeed on the object tasks without keeping the state that another question needs, which is one reason an upstream score can misjudge the benefit for a new use. Formally, consider an observation $X = (T, S)$, where T contains the supervised object-task targets and S is an additional relational state relevant to a judgment Q. If S, and therefore Q, can vary while T is fixed, then $Z = T$ and $Z' = (T, S)$ both predict T perfectly, but only Z' retains the state needed for Q; the same holds for a recognition target $T = Y$ when different occurrences share category Y. Whether a trained network keeps S is an empirical question: a predictive objective may already require the distinction, a pretrained model may retain information beyond its supervised targets, and a small probe can fail to extract what another layer makes accessible (Alain & Bengio, 2017; Hewitt & Liang, 2019; Pimentel et al., 2020; Voita & Titov, 2020; Belinkov, 2022). The question is therefore whether the state is preserved *and accessible within a stated budget*.

Many other variables could be left out in the same way. What distinguishes interaction state is that contact judgments, handover phase, state-change anticipation, and manipulation decisions depend on the same facts: who is involved, what relation holds now, and how it came about. A representation that provides participant correspondence and relational history once can serve these uses together, whereas leaving each task to re-derive them from pixels repeats the work. We hypothesize that learning participant-grounded interaction state can reduce the supervision and adaptation required for new questions about an event, not only for the questions used during training. Section 5 states the comparisons that would test this hypothesis.

Cognitive evidence motivates the target, not an architecture. Neural event patterns at multiple timescales organize naturalistic perception and extend into recall (Baldassano et al., 2017), and event organization can emerge from temporal community structure without explicit event supervision (Schapiro et al., 2013). Action decoding generalizes between videos and written descriptions only in left lateral posterior temporal cortex, while frontoparietal areas decode actions only within each type (Wurm & Caramazza, 2019), and one approximate simulation model accounts for physical judgments across several tasks (Battaglia et al., 2013). Replay follows learned abstract structure (Liu et al., 2019), shared structure and situation-specific content receive complementary codes (Samborska et al., 2022), and the Tolman–Eichenbaum Machine links shared structure to situation-

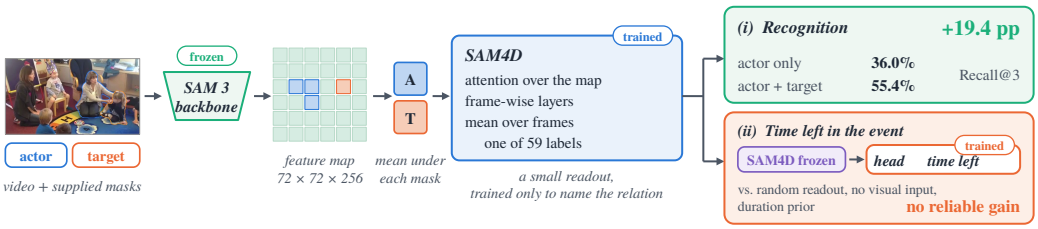

Figure 2: **The pilot pipeline.** Two supplied masks address a frozen SAM 3 feature map; SAM4D, a small readout, is trained only to name the relation and is scored on that (test i), then frozen and read by a causal head that predicts the time left in the event (test ii).

Table 1: **Pilot results.** (a) Relation recognition on 40 validation videos: Recall@1 and Recall@3 (%), the share of recorded labels ranked first or among the top three; mean ± s.d. over three training runs. (b) Predicting the time left in an event, on 15 events from eight videos: error in seconds (lower is better); Δ is P minus the row condition, so a negative value favors P.

(a) Recognition

Readout	R@1	R@3
Actor-only (U)	17.00 ± 1.14	36.04 ± 2.38
Actor–target pair (P)	34.26 ± 2.18	55.43 ± 0.58
Late fusion (L)	34.37 ± 1.68	55.99 ± 2.08
ROI pooling	35.06 ± 2.99	58.51 ± 1.89

(b) Time left in the event

Condition	Error (s)	Δ (s)
P (recognition-trained)	0.79	—
R (random readout)	0.81	−0.02
Z (zero visual)	0.76	+0.03
Duration prior (label-only)	0.61	+0.18

specific content computationally (Whittington et al., 2020). These findings suggest functional hypotheses for transferable relational organization grounded in current participants.

2.3 Make the learning priority operational

For observations $x_{\leq t}$, participating entities $\mathcal{E}$, and a downstream use k, write $z_t = f_\theta(x_{\leq t}, \mathcal{E})$ and $\hat{q}_t^{(k)} = g_{\phi_k}(z_{\leq t}, c_t^{(k)})$. The notation identifies a target, not a network: the state may be distributed across object tokens or shared context, and each use's auxiliary inputs $c_t^{(k)}$, history, and adaptation budget should be stated. Three commitments make the target concrete, and a contrast between two moments illustrates them. In two clips a giver is holding a cup; in one the receiver has not yet touched it, in the other both hands are in contact and the original grip is slackening. An object-level description—two correctly localized, tracked, and named entities and one recorded relation—is identical for both. *Grounding* ties the judgment to this cup and these two hands, including repeated occurrences involving the same pair. *Sensitivity to relational change* preserves the difference between the two moments and transitions such as release, while allowing invariance to changes the intended uses do not need; which changes count is relative to a task family. *Temporal coherence* uses the available history to distinguish a grip being established from one letting go, and supports calibrated uncertainty when hidden choices admit several futures, since visible contact shows neither whether the weight is supported nor what is intended.

Transfer is the criterion. Recognition, event-specific understanding, transfer to additional tasks or contexts within a stated adaptation budget, and physical utility in a specified decision or control outcome are complementary capabilities rather than stages to be passed in order: an effective policy need not expose an event probe, and broad transfer needs more evidence than one classifier or one temporal head. Section 5 lists candidate objectives, and the pilot of Section 3 implements one way of accessing frozen features within this framework.

3 A pilot study of recognition and interaction-state transfer

SAM4D is a small readout over frozen SAM 3 features (Carion et al., 2026) (Figure 2; Appendices A–D). From a video and the supplied masks of two participants, an actor and a target, the

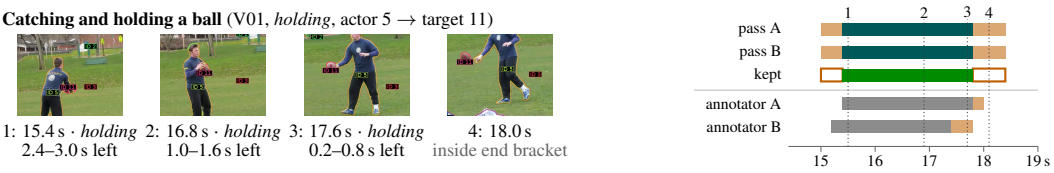

Figure 3: **Labeled case study.** One of the 15 scored events (V01, *holding*). Right: when two Gemini passes, the interval kept for scoring, and the two human annotators mark the event as active; orange marks an uncertain start or end.

Algorithm 1 Time-left evaluation of a frozen representation

Input: frozen readout f_θ, its random initialization f_{θ_0}; frame x_t, masks M_t, label y, end bracket $[\tau_t^-, \tau_t^+]$ (s)

1: $z_t^{\mathrm{P}} = f_\theta(x_t, M_t), \quad z_t^{\mathrm{R}} = f_{\theta_0}(x_t, M_t), \quad z_t^{\mathrm{Z}} = \mathbf{0}$ ▷ after the framewise layers

2: **for** $z \in \{z^{\mathrm{P}}, z^{\mathrm{R}}, z^{\mathrm{Z}}\}$; 3 video splits $\times$ 3 training runs **do**

3: $\quad \hat{T}_t = g_\phi(z_{s:t},\, y,\, t-s), \quad s = 32\lfloor t/32 \rfloor$ ▷ causal head

4: $\quad \phi^* = \arg\min_\phi \sum_t w_t\, \rho\big(\hat{T}_t, [\tau_t^-, \tau_t^+]\big)$ ▷ w_t: frame weight; $\rho = 0$ inside the bracket

5: $\hat{T}_t^{\mathrm{prior}} =$ median time left of training frames with label y ▷ duration prior

6: $d_t = \max(\tau_t^- - \hat{T}_t,\ \hat{T}_t - \tau_t^+,\ 0)$ ▷ interval error (s)

Output: $\Delta_c = \bar{d}^{\mathrm{P}} - \bar{d}^{c}$, $c \in \{\mathrm{R}, \mathrm{Z}, \mathrm{prior}\}$; $\bar{d}$: mean over runs, then frames, events, and videos

frozen SAM 3 image backbone turns each frame into a $72 \times 72 \times 256$ feature map in which each location already mixes information from the whole frame, and averaging the features under each mask gives one vector per participant. The readout—attention over the map, a framewise block that does not mix frames, mean pooling over frames, and a 59-way classifier—is trained only to name the recorded relation, the *source* task in transfer-learning usage (Zamir et al., 2018); it is then frozen, and a small causal head, one that sees only past frames, is trained on its representation to predict how much time is left in the current event. SAM4D works on 2D masks over time and builds no 3D geometry. The experiments evaluate this readout, with the participants supplied, at a stated layer, which we call the *interface* (Algorithm 1).

3.1 Observation one: specifying the participants improves recognition

The source task is relation recognition on the subset of PVSG videos that come from VidOR and have a person as actor (Shang et al., 2019; Yang et al., 2023): 2,015 training examples from 236 videos, 725 validation examples from 40 videos, and a vocabulary of 59 relation labels (*predicates*) taken from PVSG's VidOR records (Appendix A). Conditioning on the supplied actor–target pair raises Recall@3 by 19.39 percentage points over an actor-only query and approximately doubles Recall@1 (Table 1a). Late fusion scores similarly and ROI pooling slightly higher (+3.08 points), so the gain comes from access to the supplied participants rather than from one fusion mechanism. Both queries read the same feature map and therefore have the same scene context; Recall@3 measures retrieval of recorded labels rather than precision over all true relations.

3.2 Observation two: no reliable advantage in predicting the time left

Recognition asks for $P(Y \mid X)$ with Y a category; the temporal task asks for $P(T_{\mathrm{rem}} \mid X_{\le t}, Y)$, the time remaining in this occurrence given the frames so far and its category, which can change while Y stays constant (Section 2.2). Remaining time is quantitative, defined at each frame, and not determined by the label, which makes it one diagnostic of interaction understanding among several; it can be unpredictable when the end depends on choices the past frames do not show.

The 11 temporally annotated videos are removed from the source data, and the source readout is retrained once on the remaining 1,944 examples from 226 videos (actor-only, pair, and ROI Recall@3: 32.15, 55.85, 56.61%). We freeze the pair representation after its framewise layers and train a *causal* head, one that sees only past frames, to estimate remaining time on eight videos, 15 *holding* and *touching* events, and 212 event–frame pairs. Event start and end are labeled as intervals (*brackets*) by two independent passes of Gemini over each whole video (Figure 3); Section 3.4 compares these automatic labels with two human annotators. The head sees the prefix of a fixed

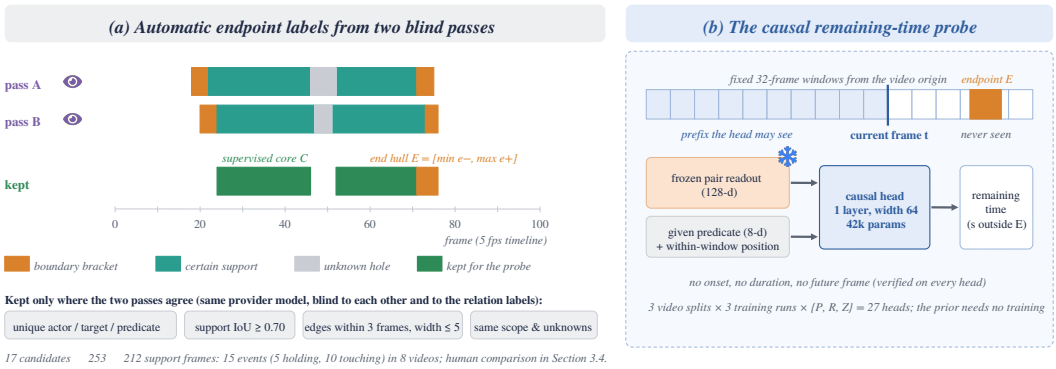

Figure 4: **Automatic labels and the causal head.** (a) Only frames on which the two Gemini passes agree are scored. (b) The head sees the prefix of a fixed 32-frame window and the given label.

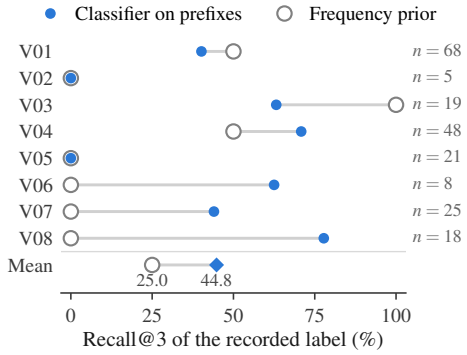

Figure 5: **Recognition from the temporal head's own prefixes.** Recall@3 of the frozen 59-way classifier per video, against the frequency prior; n is the number of scored event–frame pairs. The mean weights videos equally.

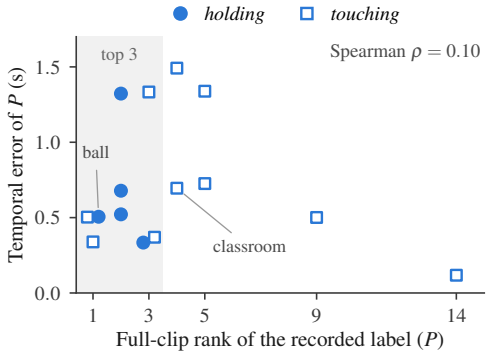

Figure 6: **Recognition rank and temporal error.** Full-clip rank of the recorded label against P's temporal error for the 15 events; shading marks the top three. The ball and classroom events are those of Figure 7.

32-frame window, the given label, and the position within the window, but no onset, duration, or future feature (Figure 4). We compare P (the recognition-trained readout), R (the same readout at its saved random initialization, over the same pretrained backbone), Z (zero visual input; label and position retained), and a *duration prior* that predicts from training-fold timing statistics of events with the same label and receives neither visual nor position input (Algorithm 1).

The interval error is the distance in seconds between the predicted end time and the labeled end bracket, zero inside it (Appendix B). Its values are P 0.7915, R 0.8106, Z 0.7580, and prior 0.6117 seconds (Table 1b). P is better than R in three of the eight videos, better than Z in four, and better than the duration prior in three, and the average differences are small next to how much they vary between videos (Figure 8). Despite a clear recognition gain, the recognition-trained representation shows no reliable advantage over the controls: in this pilot, recognition and interaction-state prediction come apart as evaluation targets. The result concerns what this head can use; whether the features hold timing information is examined in Section 3.6.

3.3 Recognition under the temporal head's own visual prefixes

Full-event recognition and the temporal head observe different inputs: a supplied clip of up to 16 sampled frames versus a fixed-window prefix. To narrow this difference we replayed the filtered source classifier at the 212 event–frame pairs, averaging the cached P features read after the frame-wise layers from the window start to the current frame and passing them to the original 59-way linear classifier (Appendix C). The replay reuses cached features without retraining, keeps the 86 prefixes

Table 2: **Gemini event labels against two human annotators.** Counts are over the 15 Gemini events used for scoring; start and end differences are medians over the nine events that both annotators record, in seconds (one frame is 0.2 s).

	Events recorded	Start difference (s)	End difference (s)
Gemini vs. annotator A	10 of 15	0.2	0.2
Gemini vs. annotator B	10 of 15	0.2	0.2
Annotator A vs. annotator B	9 of 15 by both	0.1	0.0

with missing participant observations, and uses the recorded label only for scoring. Recall@3 averaged with equal weight per video is 44.80%, against 25.00% for a prior that ranks labels by their frequency in the source training data (the *frequency prior*); the classifier is above this prior in four of the eight videos, below it in two, and equal in two (Figure 5). The recorded category is thus often retrievable from the temporal head's restricted view. The two recognition tests sample frames differently, so the difference between 44.80% and the full-clip result does not isolate the effect of history length (Appendix C).

3.4 Automatic labels compared with two human annotators

Two people annotated the same eight videos independently, without seeing the automatic labels, under shared written definitions of *holding* and *touching*; each recorded start and end ranges and marked stretches they could not judge (Appendix D). Table 2 compares each annotator's final records with the 15 automatic events used for scoring, and Figure 3 shows them for one event. Both annotators record 9 of the 15 events. On these nine, the automatic start and end times differ from each annotator's by a median of one frame (0.2 s), slightly more than the annotators differ from each other; the largest difference, 1.4 s, is at the start of one *holding* event (Table 7). The other six events are less secure: for one, both annotators record *holding* but not *touching* over the same frames; two appear only in the first version of one annotator's records; one is marked as not judgeable by both because of occlusion; and in two, one annotator records the event while the other marks it as not judgeable or records the other predicate. Restricted to the nine events both annotators record, which come from five videos, the errors are 0.55 s for P, 0.77 s for R, 0.67 s for Z, and 0.49 s for the duration prior; P is better than R in three of these five videos, and the prior remains lowest.

3.5 Case studies and video-level variation

Figure 7 presents two events with contrasting outcomes. In the ball example P ranks the recorded label first from the full clip and its temporal readout is close to the controls; in the classroom example the full-event rank is 4 while P's temporal error is below R's (0.70 versus 1.41 s); under the prefix replay their ranks range over 2–4 and 2–5. Appendices B and G show all 15 events and eight videos. Across the 15 events, P places the recorded label in its top three for 9, and its label rank and temporal error have Spearman $\rho = 0.10$, an estimate too imprecise with 15 events to say whether the two are related (Figure 6). Figure 8 shows that the aggregates hide heterogeneity, with per-video differences from about -0.7 to $+1.2$ seconds. Leaving out one video at a time from the final average, with heads, labels, and predictions fixed, can change whether P or R, and P or Z, has the lower error, while the duration prior keeps the lower error against P in all eight cases (Figure 8; Appendix C).

3.6 What the observations show and what they leave open

Two further analyses show that the timing comparison depends on which layer is read, and that a linear readout finds some remaining-time information in the trained features (Appendix E). The trained readout's error minus the random readout's error is $+0.0074$ seconds when features are read before the framewise layers and -0.0861 seconds when they are read after them. A linear regression from single-frame features read after the framewise layers, trained and tested on different videos, explains 5% of the variance in remaining time for P (17% when videos and events are weighted equally) and none for R. Specifying participants, obtaining information, and delivering a downstream benefit are thus three links, each needing its own test.

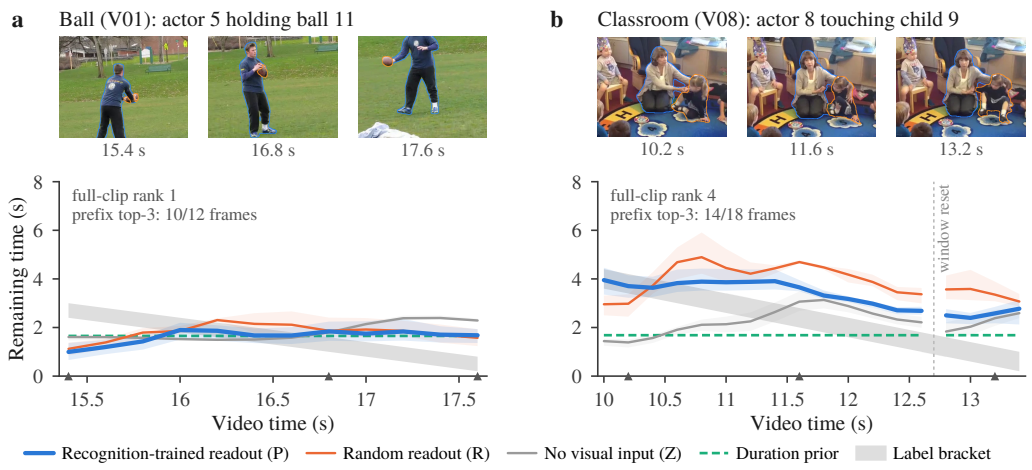

Figure 7: **Two events, two readouts.** (a) Ball, recorded *holding*; (b) classroom, recorded *touching*. Bottom: remaining time predicted by the causal head from the prefix of its fixed 32-frame window (lines, mean of three training runs; bands, range across runs; gray, labeled end interval). Full-clip ranks and prefix top-3 counts come from the recognition classifier. P's event error is 0.51 s in (a) and 0.70 s in (b), where the random readout reaches 1.41 s.

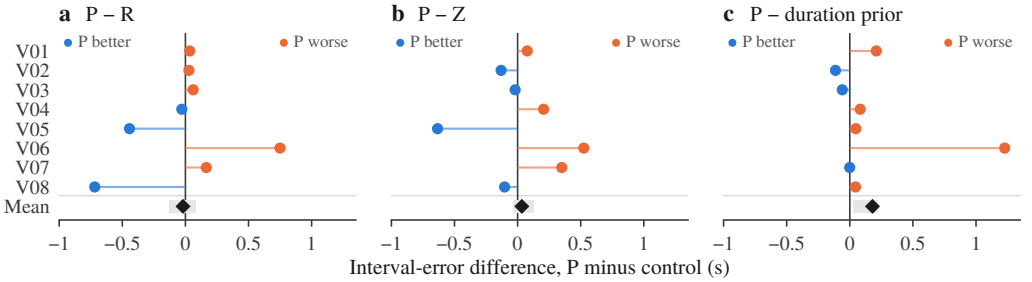

Figure 8: **The aggregate contrasts hide video-level variation.** P minus each control per video (negative favors P). Diamond: mean over the eight videos; gray band: range of this mean when one video at a time is left out.

Limitations of the pilot. The evidence comes from eight videos, enough to show that two capabilities come apart but not how large the difference is; repeated frames and training runs do not add independent samples, and the temporal analyses are several views of one sample that was explored during development. A single training run of the retrained readout, automatic labels of which six of the 15 scored events are not confirmed by both annotators, different inputs for the recognition and timing tests, and one small readout limit interpretation; a stronger head or better observations could yield a positive result. Placing the true end at other frames inside the labeled end intervals leaves the duration prior better than P on these eight videos (Appendix B); whether this holds beyond them is untested. The observations motivate the hypothesis of Section 1, that objectives which explicitly preserve relational dynamics, interaction progress, and state changes may improve event representations and their transfer; the pilot does not compare such objectives, test transfer across several tasks, or evaluate action.

4 Alternative Views

Object-centric dynamics already model interactions. Object-based representations support relational reasoning and coupled dynamics (Kipf et al., 2018; Wu et al., 2023), and scene graphs, anticipation, affordances, and state-change tasks pursue relevant targets (Ji et al., 2020; Cong et al., 2021; Peddi et al., 2024; Liu et al., 2020; Mo et al., 2021; Wu et al., 2022; Grauman et al., 2022).

These models are compatible with the position. The *more* in our title is the state of a particular interaction (Section 1); object tokens and existing backbones can carry it, and an object-centric model that exposes it meets the first part of the position. What the position adds is a test, since success on identity, localization, or persistence leaves open whether that state is preserved (Section 2.2).

General pretraining plus cheap adaptation may already do this. If large-scale predictive or generative pretraining with a small adapter or prompt answers contact, phase, and imminent-change questions as well as a targeted objective, the targeted objective only adds annotation, a fixed label vocabulary, and a pipeline to maintain (Assran et al., 2025; Zhou et al., 2025; Bardes et al., 2024; Agarwal et al., 2025). The pilot is consistent with this view, and the two parts of the position (Section 1) separate what is at stake. A pretrained model that exposes the state of a particular event meets the first part. Whether explicitly targeting that state gives better transfer or cheaper adaptation than general objectives is the second part; it is untested, and a test with comparable observations and adaptation resources would decide it (Section 5).

Task-specific learning may be preferable. A controller can exploit useful information without exposing an interpretable interaction state, and direct task training may outperform a shared representation for a narrow deployment. The position becomes practically stronger only if a shared target reduces the cost of additional tasks, improves transfer to new participants or contexts, or supports calibrated predictions that several uses can share.

A fixed label vocabulary may overconstrain learning. Contact and support labels can be useful, but labeling them exhaustively is costly and ambiguous, as the disagreements between our two annotators illustrate. Latent prediction, self-supervised temporal structure, and soft correspondence are viable routes, supported by the evidence in Section 2.2 that structure can emerge through statistical learning; the position asks for accessible relational distinctions and leaves the mechanism open.

5 How the position could be tested

Three experiments would show whether training for interaction state is worth its cost. Train the same backbone on the same data and compute with four objectives: object identity and location, interaction recognition, general prediction of video, and prediction of interaction state. Test each representation on at least two event-level questions that none was trained on, such as the time left in a contact or whether a handover has begun, with few labeled examples and on new participants and scenes. Report object-task and recognition scores alongside, so that a gain on event questions does not hide a loss elsewhere, and separate gains from extra labels, extra history, or extra capacity.

Decide what the state target should contain. A fixed list of state labels is costly and may miss states. Candidates include participant roles, contact changes, progress through an interaction, and possible next states given recent history, possibly learned by latent prediction (Bardes et al., 2024; Assran et al., 2025). Datasets that label participants, relations over time, and state changes (Shang et al., 2019; Ji et al., 2020; Yang et al., 2023; Liu et al., 2022; Grauman et al., 2022; 2024), and tasks on event boundaries and progress (Shou et al., 2021; Becattini et al., 2020), are starting points.

Find out what the pilot's features lack. Asking about a state that is visible in the current frame, such as whether two hands are in contact, once from the current frame alone and once with past frames, on new videos, would separate missing information from missing history. If existing representations answer the new questions as well and as cheaply, the first part of the position stands, but a separate objective is not needed. If a state objective helps only on the questions on which it was trained, the result supports task-specific training rather than a shared representation. Only a gain in new questions at a lower adaptation cost supports the second part.

6 Conclusion

We advocate treating the state of the interaction between particular participants, and how it changes, as an explicit learning target, compatible with object-based architectures and foundation models that already preserve such state, and judged by transfer to new event-level questions. In the pilot, specifying the participants improves recognition, while the recognition-trained representation shows no reliable advantage at predicting the time left in an event. The next step is to compare learning targets under matched conditions.

AI USE STATEMENT

Generative AI tools assisted with refining and polishing the text. They were not used to synthesize benchmark observations, ground truth, or reported experimental results. Numerical claims are checked against saved experiment records. The authors remain responsible for verifying the code, scientific claims, citations, and manuscript.

ETHICS STATEMENT

Perception of interactions can reveal behavior and relationships between people, and making event information available for additional uses can increase privacy and surveillance risks. The present pilot uses existing research video data and does not evaluate identity recognition, mental-state attribution, or safe autonomous deployment. Figures 1, 2, 3, 7, 10, 14, and 15 reproduce frames from a publicly released research video dataset (Shang et al., 2019; Yang et al., 2023); no face or identity is analyzed, and no new footage was collected. The human annotators reviewed the same released videos under a written rubric; their records are reported in aggregate and identified only by anonymous labels. Provider-generated labels can contain omissions and systematic errors. Release or further provider processing should respect the source data's permissions and applicable access restrictions. The reported temporal predictions are not validated for inferring human intent or for safety-critical decisions.

REPRODUCIBILITY STATEMENT

Appendices A–F specify the source task, retained support, frozen feature interfaces, training budgets, metrics, controls, the same-prefix replay and fixed-prediction influence analysis, the annotation audit and its comparison with the automatic labels, and interpretation limits. The accompanying writing package contains machine-readable aggregate results, held-out predictions, per-event recognition ranks, the replay protocol and outputs, the label-comparison script and per-event outputs, and frozen features underlying Figures 5, 6, 7, 8, 11, and 13, together with scripts for regenerating the result tables and data plots. It does not contain the original feature caches, checkpoints, or videos required to reproduce model training, which require authorized access. The matched learning-objective comparisons, matched-observation diagnoses, and multi-task, cross-context, and physical-utility studies proposed in Section 5 remain future work.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical AI. *arXiv:2501.03575*, 2025.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*, 2017.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv:2506.09985*, 2025.

Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *CVPR*, 2023.

Christopher Baldassano, Janice Chen, Asieh Zadbood, Jonathan W Pillow, Uri Hasson, and Kenneth A Norman. Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–721.e5, 2017. doi: 10.1016/j.neuron.2017.06.041.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *TMLR*, 2024.

Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *PNAS*, 110(45):18327–18332, 2013. doi: 10.1073/pnas.1306572110.

Federico Becattini, Tiberio Uricchio, Lorenzo Seidenari, Lamberto Ballan, and Alberto Del Bimbo. Am I done? Predicting action progress in videos. *ACM TOMM*, 16(4):1–24, 2020.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Comput. Linguist.*, 48:207–219, 2022.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. In *RSS*, 2025. doi: 10.15607/RSS.2025.XXI.010.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL*, volume 229 of *PMLR*, pp. 2165–2183, 2023.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *ICLR*, 2026.

Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018.

Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *ICCV*, 2021.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *IJCV*, 130 (1):33–55, 2022.

Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How well do self-supervised models transfer? In *CVPR*, pp. 5414–5423, 2021.

Antonino Furnari and Giovanni Maria Farinella. What would you expect? Anticipating egocentric actions with rolling-unrolling LSTMs and modality attention. In *ICCV*, 2019.

Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *ICCV*, 2021.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *CVPR*, 2024.

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv:2012.05208*, 2020.

Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. ConceptGraphs: Open-vocabulary 3D scene graphs for perception and planning. In *ICRA*. IEEE, 2024.

Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv:1505.04474*, 2015.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *EMNLP-IJCNLP*, pp. 2733–2743, 2019.

Yingdong Hu, Renhao Wang, Li Erran Li, and Yang Gao. For pre-trained vision models in motor control, not all policy learning methods are created equal. In *ICML*, volume 202 of *PMLR*, pp. 13628–13651, 2023.

Nathan Hughes, Yun Chang, and Luca Carlone. Hydra: A real-time spatial perception system for 3D scene graph construction and optimization. In *RSS*, 2022.

Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*, 2020.

Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. HOTR: End-to-end human-object interaction detection with transformers. In *CVPR*, pp. 74–83, 2021.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, et al. OpenVLA: An open-source vision-language-action model. In *CoRL*, volume 270 of *PMLR*, pp. 2679–2713, 2025.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *ICML*, volume 80 of *PMLR*, pp. 2688–2697, 2018.

Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. In *ICLR*, 2022.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better ImageNet models transfer better? In *CVPR*, 2019.

Christopher A. Kurby and Jeffrey M. Zacks. Segmentation in the perception and memory of events. *Trends Cogn. Sci.*, 12:72–79, 2008.

Miao Liu, Siyu Tang, Yin Li, and James M. Rehg. Forecasting human-object interaction: Joint prediction of motor attention and actions in first person video. In *ECCV*, 2020.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4D egocentric dataset for category-level human-object interaction. In *CVPR*, 2022.

Yunzhe Liu, Raymond J Dolan, Zeb Kurth-Nelson, and Timothy E J Behrens. Human replay spontaneously reorganizes experience. *Cell*, 178(3):640–652.e14, 2019. doi: 10.1016/j.cell.2019.06.012.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with Slot Attention. In *NeurIPS*, volume 33, 2020.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an artificial visual cortex for embodied intelligence? In *NeurIPS*, volume 36, 2023.

Kaichun Mo, Leonidas Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2Act: From pixels to actions for articulated 3D objects. In *ICCV*, 2021.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. *arXiv:2603.14482*, 2026.

Rohith Peddi, Saksham Singh, Saurabh, Parag Singla, and Vibhav Gogate. Towards scene graph anticipation. In *ECCV*, 2024.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *ACL*, pp. 4609–4622, 2020.

Veronika Samborska, James L Butler, Mark E Walton, Timothy E J Behrens, and Thomas Akam. Complementary task representations in hippocampus and prefrontal cortex for generalizing the structure of problems. *Nat. Neurosci.*, 25(10):1314–1326, 2022. doi: 10.1038/s41593-022-01149-8.

Anna C Schapiro, Timothy T Rogers, Natalia I Cordova, Nicholas B Turk-Browne, and Matthew M Botvinick. Neural representations of events arise from temporal community structure. *Nat. Neurosci.*, 16(4):486–492, 2013. doi: 10.1038/nn.3331.

Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *ACM MM*, pp. 1300–1308, 2017.

Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *ICMR*, pp. 279–287, 2019.

Mike Zheng Shou, Stan Weixian Lei, Weiyao Wang, Deepti Ghadiyaram, and Matt Feiszli. Generic event boundary detection: A benchmark for event segmentation. In *ICCV*, 2021.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *EMNLP*, pp. 183–196, 2020.

James C R Whittington, Timothy H Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy E J Behrens. The tolman-eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263.e23, 2020. doi: 10.1016/j.cell.2020.10.024.

Ruihai Wu, Yan Zhao, Kaichun Mo, Zizheng Guo, Yian Wang, Tianhao Wu, Qingnan Fan, Xuelin Chen, Leonidas Guibas, and Hao Dong. VAT-Mart: Learning visual action trajectory proposals for manipulating 3D ARTiculated objects. In *ICLR*, 2022.

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

Moritz F Wurm and Alfonso Caramazza. Distinct roles of temporal and frontoparietal cortex in representing actions across vision and language. *Nat. Commun.*, 10(1):289, 2019. doi: 10.1038/s41467-018-08084-y.

Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, and Ziwei Liu. Panoptic video scene graph generation. In *CVPR*, 2023.

Jeffrey M. Zacks and Khena M. Swallow. Event segmentation. *Curr. Dir. Psychol. Sci.*, 16:80–84, 2007.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, pp. 3712–3722, 2018.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *ICML*, volume 267 of *PMLR*, pp. 79115–79135, 2025.

A Source recognition: protocol and participant conditioning

SAM 3 combines concept-prompted detection and segmentation with tracking (Carion et al., 2026). Its shared vision encoder receives image–text pretraining and image-level detection/segmentation training; it is frozen during final tracker training. The present feature study therefore does not isolate the effect of tracking supervision.

Table 1a evaluates how participant conditioning affects relation recognition, and Table 3 adds macro-averaged recall; Figure 2 shows what is supplied, frozen, trained, and tested, and Figure 9 places this source study within the temporal evaluation protocol. U denotes the actor-only query, P the supplied actor–target pair query, L the pair with late fusion, and ROI participant ROI pooling.

Table 3: **Full source recognition ablation.** Recall of the recorded relation labels (%), mean ± s.d. over three source seeds; participants and event intervals are supplied. Pair conditioning raises R@3 by 19.39 percentage points over actor-only conditioning.

Readout	R@1	R@3	Macro R@3
Actor-only query (U)	17.00 ± 1.14	36.04 ± 2.38	18.59 ± 1.35
Actor–target pair query (P)	34.26 ± 2.18	55.43 ± 0.58	36.26 ± 0.90
Pair, late fusion (L)	34.37 ± 1.68	55.99 ± 2.08	33.90 ± 3.45
Participant ROI pooling (ROI)	35.06 ± 2.99	58.51 ± 1.89	36.09 ± 1.47

Data and observation contract. We use the human-actor slice of the VidOR-sourced portion of PVSG (Shang et al., 2019; Yang et al., 2023). Each example supplies an actor track, target track, relation interval, and nonempty recorded predicate set. Each interval is a maximal contiguous segment with a constant recorded label set; its boundaries have not been independently certified as physical event boundaries. Up to 16 native frames are sampled uniformly within it. The video-disjoint source split contains 2,015 examples from 236 training videos and 725 examples from 40 validation videos, with 2,202 and 899 positive assignments, respectively. The 59-class output vocabulary follows relation strings in the frozen VidOR-sourced records, not the top-level PVSG list or the original VidOR benchmark vocabulary: the released PVSG annotation lists 57 relation classes (Yang et al., 2023), while the relation records of its 289 VidOR-sourced videos use 59 distinct strings, 54 from that list and five that are not on it (climbing, enclosing, getting down on, going down, squatting on). Training and validation support 49 and 47 predicates; six validation predicates lack training positives and remain in the evaluation.



Figure 9: **What is trained, what is frozen, and what is tested.** The recognition study (top), the temporal readout interface (middle), and the four conditions of the remaining-time probe (bottom). Participant masks select locations in a shared frozen SAM 3 feature map. Condition names: U actor-only, P actor–target pair, L late fusion, ROI participant pooling.

Shared features and participant conditioning. The frozen SAM 3 image backbone (Carion et al., 2026) produces a $72\times72\times256$ feature map, the third of the four outputs of the feature pyramid on top of its vision transformer; each location of this map already integrates context from the whole frame. Frames are resized to 1008×1008, normalized with mean and standard deviation 0.5, encoded in float32, and cached in float16. Masks address participant queries within this shared feature map. The input is the visual feature tensor rather than a text-conditioned segmentation output; descriptors may therefore contain information beyond the masked participant. Tracks and masks are supplied, so the study evaluates the backbone features rather than the tracking system, and these features can carry more than object identity and location.

Let $\mathcal{P}(M)$ denote patches touched by mask M. A participant seed is

$$\mu(F_t, M) = |\mathcal{P}(M)|^{-1} \sum_{p\in\mathcal{P}(M)} WF_{t,p}. \tag{1}$$

Pooling weights occupied patches equally, not proportionally to mask area. Three learned interaction/actor/target slots receive role embeddings; participant slots additionally receive their mask-pooled seeds. Missing support uses a learned role-specific token. Two cross-/slot-attention layers use width 128, eight heads, and dropout 0.1. Their three output slots are concatenated, normalized, projected to 128 dimensions, and passed through GELU. P allows slot exchange; L disables it while retaining field attention and a shared zero key/value anchor. U replaces the target-conditioned seed with an input-independent slot, exposing neither target mask support nor target missingness. ROI uses participant pooling and a roughly parameter-matched MLP instead of readout attention. No condition receives explicit entity categories or handcrafted geometry. U retains access to the feature map, which may contain target information, so the U–P contrast tests explicit participant conditioning rather than the removal of target evidence.

Source objective and readout interface. After the 128-dimensional readout, the classifier path applies two Transformer layers with sinusoidal positions and diagonal-only, *framewise* attention, then averages valid outputs and predicts 59 scores. This block does not exchange information across frames. For scores s and recorded positives Y^+, training minimizes

$$\mathcal{L}_{\text{rel}} = -\frac{1}{|Y^+|} \sum_{c\in Y^+} \log \frac{\exp(s_c)}{\sum_{j=1}^{59} \exp(s_j)}. \tag{2}$$

Each sample has equal weight. Unrecorded labels compete in the denominator; they are not certified negatives and do not have zero gradient. The objective is a ranking surrogate and does not yield calibrated multi-label probabilities. AdamW uses learning rate 10^{-4}, weight decay 0.01, batch size eight, gradient clipping at 1.0, and ten epochs. The three seeds are 42, 123, and 456; fixed final checkpoints are reported without validation-based selection.

Metric and uncertainty. Recorded-positive recall averages $|Y_i^+ \cap \text{Top}_k(s_i)|/|Y_i^+|$ equally over examples. Cutoff ties receive their expected inclusion credit rather than arbitrary class-index tie breaking. The main table reports sample standard deviations across the three training runs. Contrast intervals use 2,000 paired resamples of the 40 validation videos, seed 20260908, retaining within-video dependence and averaging the three fitted-seed metrics. They exclude model retraining, are unadjusted for multiple comparisons, and use a validation set already used in development. The 95% interval of the Recall@3 difference is $[14.29, 24.95]$ points for the pair query minus the actor-only query and $[0.62, 5.63]$ for ROI pooling minus the pair query.

Separation for the temporal evaluation. The temporal study uses a separately retrained source model to keep its evaluation videos outside source training. The 11 completed temporal-annotation videos are excluded from source training and validation. Removing the ten with source-training examples leaves 1,944 examples from 226 videos. Retraining with source seed 42 gives R@3 of 32.1494%, 55.8506%, and 56.6092% for U, P, and ROI. The random readout (R) is the complete paired source model saved before its first optimization step, with the same pretrained, frozen SAM 3 image backbone; only the readout is at its random initialization.

B EVENT-SPECIFIC EVALUATION: REMAINING TIME AND ENDPOINT AUDIT

Table 1b and Table 4 evaluate remaining-time prediction and its sensitivity to uncertain endpoints. Algorithm 1 and Figure 4 (Section 3.2) summarize the procedure, the automatic labels, and the

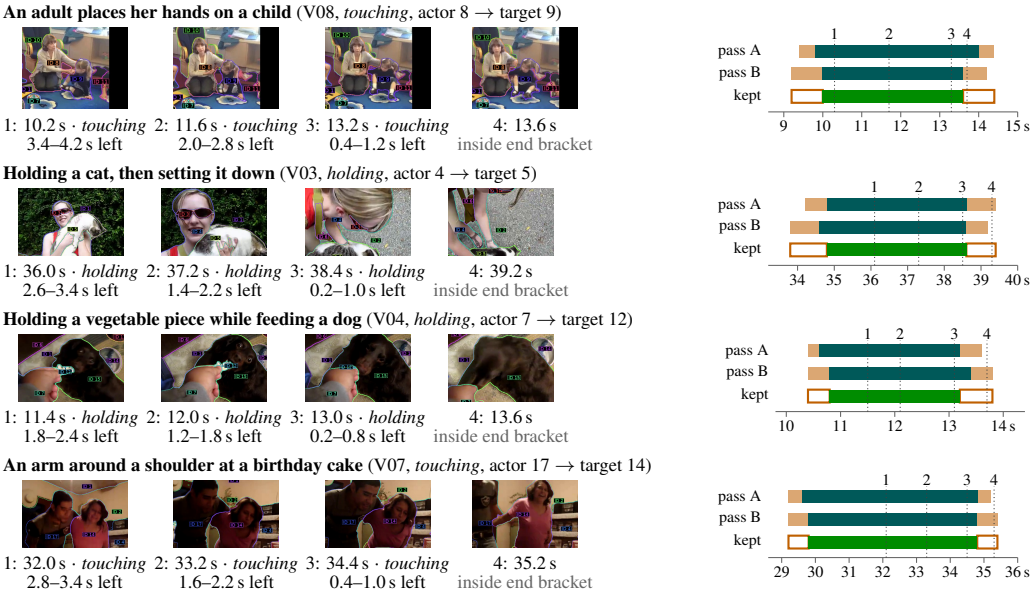

Figure 10: **Four additional evaluated events.** Encoding follows Figure 3, without the human rows. Frames 1–3 lie in retained support, with remaining time derived from the endpoint bracket; frame 4 lies within that bracket. Disagreements between annotation passes (V03 onset, V04 end) narrow the retained core; endpoint targets use the conservative hull. These whole-video Gemini proposals supply the automatic labels. Video-level errors (P/R/Z/prior, seconds) are V08 0.70/1.41/0.80/0.65, V03 0.68/0.62/0.70/0.74, V04 0.42/0.45/0.22/0.34, and V07 1.49/1.33/1.14/1.49.

Table 4: **On this sample, the prior keeps its advantage at each admissible endpoint.** Ranges of point-endpoint MAE differences (s) over endpoints shared across conditions within each event's bracket, with predictions fixed. These are deterministic sensitivity ranges, not confidence intervals.

Comparison	Admissible difference range	Can the sign change?
P minus R	$[-0.0868, +0.0483]$	Yes
P minus Z	$[-0.0761, +0.1474]$	Yes
P minus label-only prior	$[+0.0168, +0.3426]$	No; P worse

inputs of the causal head. Figures 3 and 10 show five of the fifteen evaluated events with both automatic annotation passes; Figure 11 shows the held-out predictions for all fifteen.

Automatic labels and retained support. Two whole-video Gemini annotation passes (recorded model identifier `gemini-3.8-flash`, maximum 16,384 output tokens) inspect synchronized RGB and mask-contour/track-ID panels, a track roster, and eight operational predicate definitions. Each pass is blind to the other's outputs and source relation labels. Both use the same provider model, so their agreement measures consistency and can retain shared errors; Section 3.4 and Appendix D compare the resulting automatic labels with two human annotators. The displayed timeline is 5 fps. Missing events, occlusions, unknown intervals, and absent masks are treated as unknown, not as negatives. Gemini serves only as the label generator: its passes inspect the whole video, whereas the evaluated causal heads receive a fixed-window prefix.

Throughout, an event's *support* is the set of frames on which it is labeled as active. Candidates require unique same-actor/target/predicate matches, at least 0.70 certain-support IoU, corresponding bracket coordinates within three frames, per-pass bracket widths at most five frames, and agreement on scope, censoring, and unknown holes. Only finite, uncensored, positive observed events enter the probe. With onset brackets $S^j = [s^j_-, s^j_+]$, end brackets $E^j = [e^j_-, e^j_+]$, and unknown regions $\mathcal{U}^j$

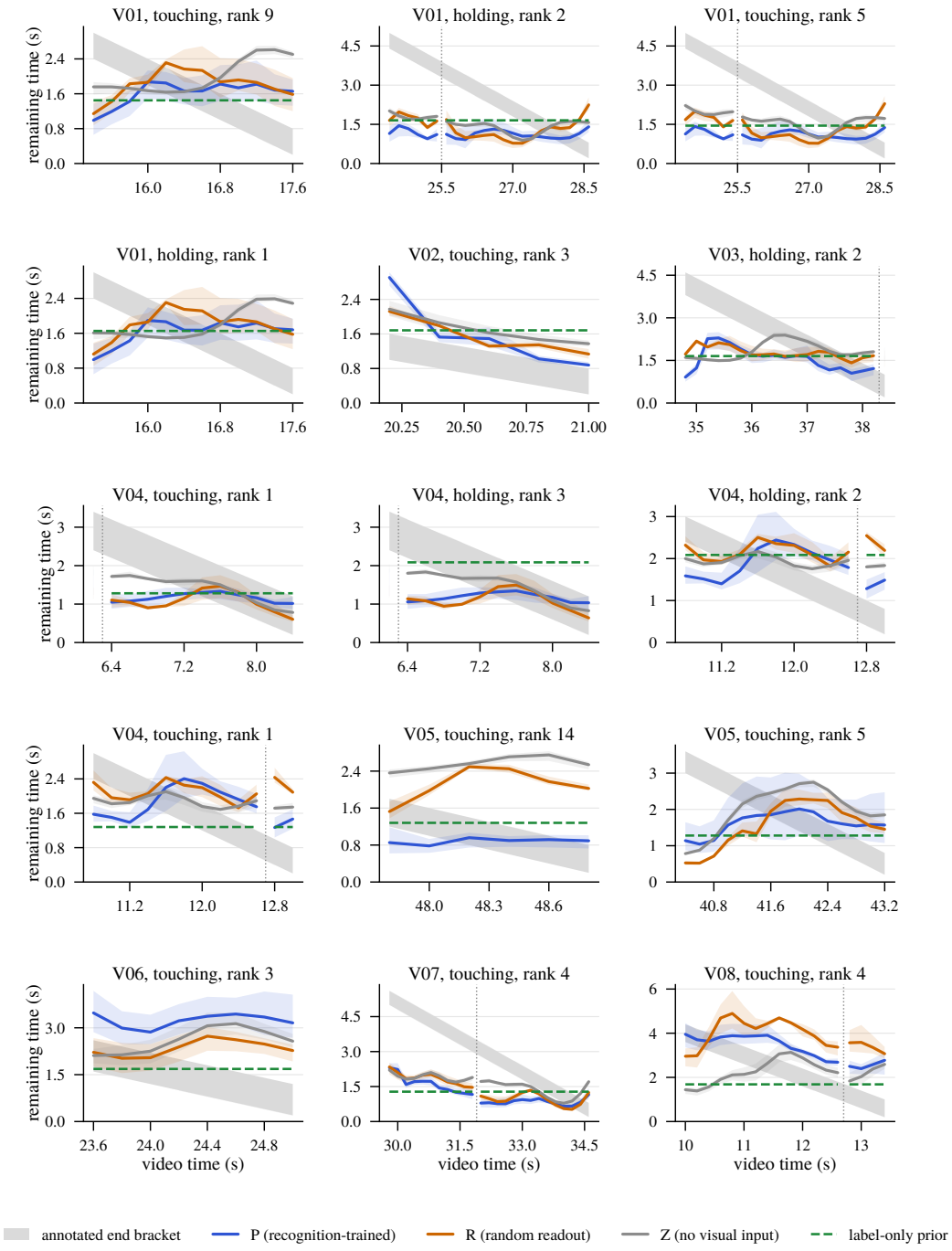

Figure 11: **Held-out predictions on all 15 evaluated events.** Lines are seed means, bands seed ranges, grey regions the automatic-label end brackets, and the dashed line the label-only prior. Titles show the video, recorded predicate, and its full-clip rank under P. Lines break at missing support and 32-frame window resets (dotted). Across 212 support contexts, the seed-mean predictions have small descriptive correlations with remaining-time bracket midpoints (Pearson $r = 0.05$, 0.04, -0.05 for P, R, Z).

from passes $j = A, B$, the supervised core is

$$\mathcal{C} = \Big[\max_j s_+^j,\ \min_j e_-^j\Big) \setminus \bigcup_j \mathcal{U}^j. \tag{3}$$

Boundary targets retain the conservative hull across passes, including $E = [\min_j e_-^j, \max_j e_+^j]$. Contradictions and unavailable current participant masks further remove support; local checks do not certify whole-event boundaries. Seventeen candidates provide 253 positive support frames. Current-mask exclusions remove 28 frames, and test predicates unseen in their training fold remove another 13. Evaluation therefore covers 212 event–frame contexts from 15 events—five *holding* and ten *touching*—in eight videos. These contexts correspond to 154 distinct participant-pair/frame feature keys; simultaneous predicates can share visual evidence. The other three completed videos have no eligible supervision, but remain excluded from source training.

Frozen interface and access. Folds are assigned by sorting all 11 video IDs by SHA256 and distributing them round-robin, without label-aware reassignment. Each eligible video is held out once. Every native frame is encoded; participant pairs are deduplicated without using the predicate to compute their representation. The original equal-occupied-patch query mapping is preserved. The study uses POST: the paired 128-dimensional readout after the frozen source framewise block, before source mean pooling and classification. Its sinusoidal positions reset in fixed 16-frame blocks. Dense per-frame use changes the observation conditions from the source model's supplied intervals and uniform clip sampling; preserving the block does not preserve its source input distribution.

The downstream head receives the layer-normalized representation and an eight-dimensional given-predicate indicator. One width-64 Transformer layer has four heads, feed-forward width 128, dropout 0.1, and one softplus output: 42,305 parameters. Fixed windows start every 32 frames from the video origin. A strict causal mask permits only the window start through the current frame, not a sliding full-video history. Within-window positions are available; event identity, onset, duration, boundaries, and future features are not. The no-visual control zeros the 128-dimensional input but retains the same predicate and position inputs. Evaluation conditions on supplied participants, predicate, and active-event support. It measures conditional remaining-time prediction, without evaluating event detection or the online acquisition of the supplied tracks.

Training and evaluation units. For positive frame f, rate $r = 5$, and exclusive-end bracket $[e_-, e_+]$, the log-distance target is

$$B_f = \left[\log\Big(1 + \frac{e_- - f}{r}\Big),\ \log\Big(1 + \frac{e_+ - f}{r}\Big)\right]. \tag{4}$$

For predicted log-distance $\hat{\ell}_f$, the weighted loss is $\sum_f w_f\, \mathrm{SmoothL1}\big(\hat{\ell}_f, \Pi_{B_f}(\hat{\ell}_f)\big)$, with zero loss inside the bracket and for unknown rows. Weights balance frames within events, events within videos, and videos. AdamW uses learning rate 10^{-3}, weight decay 0.01, batch size eight, clipping at 1.0, and fixed epoch 40. Three conditions, three folds, and probe seeds 42/123/456 yield 27 trained heads, without source fine-tuning or checkpoint selection.

Evaluation inverts the log transform, $\hat{e}_f = f + r(\exp \hat{\ell}_f - 1)$, and measures error in *seconds*, not log units:

$$d_f = \frac{\max(e_- - \hat{e}_f,\ \hat{e}_f - e_+,\ 0)}{r}. \tag{5}$$

The label-only prior is the training-fold weighted median of $\frac{1}{2}(\inf B_f + \sup B_f)$ per predicate, with a pooled fallback. It receives the given predicate but no visual features, within-window position, labeled start time, or test duration. This differs from the no-visual learned head, which retains both predicate and position, and from the frequency prior for classification of Appendix C. We average probe seeds, frames within events, events within videos, then videos equally. Two thousand paired video-bootstrap draws (seed 20260917) provide model-conditional intervals, without multiplicity correction, source-seed uncertainty, or annotation-error coverage. For P minus R, Z, and the duration prior, the 95% intervals are $[-0.30, 0.25]$, $[-0.20, 0.24]$, and $[-0.02, 0.49]$ seconds; all include zero. Videos are the independent evaluation units; neither probe seeds nor repeated event–frame contexts increase the sample beyond eight videos. The data and POST interface were explored earlier, so this test is exploratory rather than confirmatory. Future-feature replacement, future-validity changes, and explicit prefix truncation pass for all 27 heads, with maximum log-distance deviation 2.3842×10^{-7}.

Table 5: **Same-prefix recognition replay.** The frozen 59-way source classifier applied to the temporal head's cached visual prefixes for all 212 contexts, 15 events, and eight videos; recorded-predicate Recall@3 averaged over frames within events, events within videos, then videos equally. The paired P-minus-prior contrast is $+19.80$ percentage points, 95% video bootstrap $[-4.63, 45.40]$. The frequency prior is a classification baseline distinct from the duration prior of Table 1b.

Condition	Video-balanced R@3 (%)	95% video bootstrap
Frozen P classifier	44.80	$[23.35, 63.56]$
Frequency prior (source label counts)	25.00	—

Common-endpoint audit. The prior has the lowest observed interval error. To test whether resolving uncertain boundaries could reverse its ordering against P, we fix all predictions and support after training and enumerate every integer endpoint $b_e \in [e_-, e_+]$ for each event. A candidate is shared by all conditions, frames, and seeds for that event. For each comparison we average $|\hat{e}_f - b_e|/r$ over frame/seed predictions before event/video aggregation, then independently choose admissible event endpoints to obtain the minimum and maximum aggregate difference (Table 4). Differences against R and Z can change sign; the difference against the prior keeps its sign, with P worse by 0.0168–0.3426 seconds. The resulting range is a deterministic sensitivity range for *point-endpoint MAE*, not a confidence interval for Equation 5. It adds no endpoint candidates outside the original brackets or cross-event ordering constraints. This robustness holds within the fixed sample; the corresponding interval-error bootstrap comparison includes zero. The end-only causal metric and the earlier bidirectional onset/end averages measure different quantities, so their difference does not isolate an effect of causal access.

Held-out predictions and per-event recognition. Figures 7 and 11 show the 2,120 held-out predictions of the 27 heads and the prior on the 212 support contexts. Each head predicts only its test fold. Curves show the seed mean and bands the seed range, which is not a confidence interval; the error reported in Table 1b is computed per prediction, not from the mean curve. For the per-event recognition ranks, the retrained source readouts (P, U, ROI) scored each event's full annotated clip with the original sampler (up to 16 uniformly spaced frames, clip-local positions, cached PRE readouts promoted from float16); the rank is that of the recorded predicate among the 59 scores, with no ties among target scores. This measures retrieval of the recorded predicate, not precision over all true interactions. P, U, and ROI place the recorded predicate in their top three for 9, 5, and 8 of the 15 events. Spearman's ρ between P's rank and P's per-event temporal error (frames equally weighted within each seed, then seeds) is 0.1049 with an 8-video block-bootstrap interval of $[-0.4734, 0.6621]$ (2,000 resamples); this interval is too wide to show whether the two are related. Recognition observes the supplied full event clip, whereas the temporal head observes only a fixed-window prefix; Appendix C reports the replay that narrows this difference. Both the task and the readout interface change, so this comparison does not isolate an effect of recognition training on information content.

C Same-prefix recognition replay and fixed-prediction influence analysis

Both analyses in this appendix use saved models, features, labels, and predictions and leave the timing scores unchanged; the replay protocol was written before the replay was run.

Replay protocol. For each of the 212 temporal support contexts, the cached float32 P POST vectors of every native frame from the fixed 32-frame window start through the current frame, inclusive, are averaged and passed to the original 59-way linear classifier ($[59 \times 128]$ weights and bias) extracted from the verified filtered source checkpoint used for the temporal study (Appendix A). No future frame, onset input, supplied predicate, extra normalization, or repeated source framewise block enters the computation. Frames on which a participant is missing contribute their cached representations rather than being dropped or padded; all 86 prefixes containing such frames and all 115 prefixes crossing a source 16-frame positional reset are retained; no context is excluded. Ranks are taken over the full 59-class vocabulary in descending logit order with exact ties broken by ascending

Table 6: **Human annotation audit.** Predicate-specific event records and uncertainty records of two annotators over nine participant-pair tasks in the eight evaluation videos. A-v2 is annotator A's self-review revision, made without seeing B-v1.

Version	Holding	Touching	Unknown	Unresolved	Right-censored
A-v1 (independent initial)	5	7	3	0	0
B-v1 (independent initial)	10	10	2	0	2
A-v2 (A's self-review revision)	10	10	2	1	0

vocabulary index; the recorded predicate is used only for scoring. Because the POST feature does not receive the predicate, the 212 contexts correspond to 154 distinct physical prefixes, and the 58 prefixes scored under two recorded predicates produce identical logits.

Scores and baseline. The primary score averages binary recorded-predicate Recall@3 over frames within an event, events within a video, and then the eight videos equally, matching the temporal aggregation (Table 5). The result is 44.80% (per video 40.15, 0.00, 63.16, 70.83, 0.00, 62.50, 44.00, and 77.78% for V01–V08, which contribute 68, 5, 19, 48, 21, 8, 25, and 18 contexts); the context-pooled hit count is 103/212 (48.58%). The frequency prior ranks predicates by their recorded-positive counts over the 1,944 filtered source-training samples in 226 videos (top three: *holding*, *sitting on*, *looking at*), using no timing-set labels; its video-balanced score is 25.00%. Two thousand video bootstrap resamples (`numpy.random.default_rng`, seed 20260925, percentile 2.5/97.5) give $[23.35, 63.56]\%$ for P and $[-4.63, +45.40]$ percentage points for the paired P-minus-prior difference of $+19.80$. These intervals condition on one fitted source classifier and the historical automatic labels; they do not cover source retraining or annotation uncertainty.

Verification. The checkpoint SHA-256, POST caches, vocabulary, and bound source code were matched before execution. All 212 target ranks and top-three orders agree between the float32 path, an independent float64 path, and scalar `math.fsum` checks (error below 10^{-12}). The 15 full-event ranks were recomputed from archived full-clip logits and matched. The replay reproduces the original linear operation algebraically.

Interpretation. The replay measures how much of the recorded category the frozen recognition interface can retrieve from the same cached visual prefixes that the temporal head receives. The two tasks still differ in difficulty: the temporal head is given the predicate and window position while the classifier must predict the predicate, and the historical 9/15 full-event count uses sparse clip sampling with clip-local positions whereas the replay uses dense POST features with positions reset every 16 frames. The difference between the two scores therefore does not isolate history length, training objective, or backbone information.

Fixed-prediction video influence. Starting from the saved per-seed predictions and scored support, the four headline interval errors of Table 1b were recomputed with discrepancies below 10^{-12} seconds. Each of the eight evaluation videos was then omitted in turn from the final equal-video average only, keeping every head, fold, prior, label, and prediction fixed. Over the eight omissions, $\mathrm{P} - \mathrm{R}$ ranges over $[-0.1289, +0.0808]$ seconds, $\mathrm{P} - \mathrm{Z}$ over $[-0.0366, +0.1289]$, and P minus the duration prior over $[+0.0301, +0.2216]$. The signs of the R and Z contrasts depend on which video is removed; the prior contrast stays positive in all eight omissions, while its full eight-video bootstrap interval includes zero. These ranges describe how individual videos influence the point estimates; they involve no retraining and are not confidence intervals.

D HUMAN ANNOTATION AUDIT: RUBRIC, VERSIONS, AND COMPARISON

The audit collects independent human records of the events the pilot scores. It makes the uncertainty in the automatic labels concrete and reports coverage and disagreement, which bound how the pilot's scores should be read.

Tasks and materials. Nine participant-pair tasks cover the eight evaluation videos and the nine actor–target pairs of the temporal pilot; one video contributes two pairs and appears in two tasks. Each task supplies the full 5 fps video with muted playback, frame-by-frame stepping, lossless

single-frame display, and optional supplied-track contours for the actor (blue) and target (orange) used only to confirm identity. Annotators do not see the automatic labels, previous boundaries, paper case descriptions, model scores, or prediction curves. Frame indices are integers from zero; an event is recorded as $[s, e)$ with an inclusive start candidate range $[s_-, s_+]$ and an inclusive range $[e_-, e_+]$ for the first frame no longer in the event, so that only boundaries the annotator can actually delimit receive equal lower and upper values.

Rubric. The operational definitions keep the semantics of the automatic protocol. *Holding* requires visible support or control of the target by a body part of the actor (a hand, arm, mouth, or other), not proximity, a hand resting on a surface, brief contact, or two-dimensional contour overlap; it ends when visible control is lost, not when the target leaves the frame or the mask disappears. *Touching* requires visible evidence of direct body or surface contact between the supplied participants, excluding perspective overlap, mask overlap, and contact made only through a held tool. Either predicate may be recorded without the other, both may hold simultaneously with different boundaries, and the same predicate may recur as separate events with distinct local identifiers. Annotators record *unknown* intervals $[u_-, u_+)$ with a reason (occlusion, out of view, blur, small target, ambiguous identity, mask misalignment, other) instead of guessing; an unknown interval is neither a positive nor a negative. A left- or right-*censored* record states that the relation is directly visible at the first or last source frame. An *unresolved* boundary states that a visible segment exists but one side cannot be delimited for a stated reason; its boundary values are null, which does not mean that the relation continued to the end of the video. Marking a predicate as not observed does not certify its absence throughout the task, and unknown intervals are still recorded.

Independence and versions. Two annotators (anonymous labels A and B) each completed all nine tasks and both predicate checks per task, exporting their files before any comparison. The coordinating author confirmed that the initial submissions (A-v1, B-v1) were made independently, without discussing specific answers, and that A's later revision (A-v2) followed only A's own self-review notes, without access to B's records or the pairwise comparison. A-v2 is therefore a second version of one annotator's judgment, not a third independent annotation. All three exports, the comparison reports, and validation results are frozen with byte counts and SHA-256 digests. Table 6 lists their event counts: A-v1 recorded 12 events (five *holding*, seven *touching*) with three unknown intervals; B-v1 recorded 20 (ten and ten) with two unknown intervals and two right-censored records; A-v2 recorded 20 (ten and ten) with two unknown intervals and one unresolved endpoint. Human event counts and the pilot's 15 scored events are different denominators; the comparison below maps human records onto the 15 events.

Mechanical comparison. Candidate correspondences pair records with the same task and predicate whose possible activity ranges overlap. The initial comparison (A-v1 versus B-v1) found eight isolated candidate correspondences, four records only in A-v1, twelve only in B-v1, no split/merge component, and 58 disagreeing field rows across the paired records. The revised comparison (A-v2 versus B-v1) found 15 candidate correspondences, two records only in A-v2, three only in B-v1, one split/merge component in task 05, and 63 disagreeing field rows. These counts describe the records; they are not agreement rates or accuracy measures. The disagreements concern participant identity under occlusion, contact versus control, endpoints when visibility is lost, and split versus merged occurrences; they are retained as recorded.

Comparison with the automatic labels. Section 3.4 compares the annotators' records with the 15 automatic events used for scoring (Table 7). The rules were fixed before the summary was computed and follow the comparison between annotators above. An automatic event and a human record correspond when they concern the same participants and predicate and their possible activity ranges overlap; the automatic range is the hull of the two Gemini passes, which is also the scoring target, and if several human records overlap one event, the one with the largest overlap is used. An event without a corresponding record is classified as recorded under the other predicate, covered by a stretch the annotator marked as not judgeable, or not recorded. Boundary differences are the distances in seconds between the midpoints of the start ranges and of the end ranges; censored or unresolved boundaries are excluded. The main comparison uses each annotator's final records (A-v2 and B-v1). A-v1 records 7 of the 15 events, including the two events at 21–29 s in V01 that neither final version contains. Restricting the timing evaluation to the nine events that both final versions record keeps the original predictions, labels, and scores and selects events by the human records alone. The analysis script and per-event outputs are part of the accompanying package.

Table 7: **Each scored event against each annotator version.** Automatic start and end ranges in seconds (hull of the two Gemini passes). For a recorded event, the cell gives the start and end differences in seconds between midpoints; otherwise it states whether the annotator recorded the other predicate over the same frames, marked the stretch as not judgeable, or recorded nothing.

Video	Predicate	Start (s)	End (s)	A (final, A-v2)	B (B-v1)	A (first, A-v1)
V01	*holding*	15.0–15.4	17.8–18.4	0.2 / 0.2	0.0 / 0.5	0.2 / 0.1
V01	*touching*	15.0–15.4	17.8–18.4	other predicate	other predicate	other predicate
V01	*holding*	21.2–21.6	28.8–29.4	none	none	0.1 / 0.1
V01	*touching*	21.2–21.6	28.8–29.4	none	none	6.1 / 0.6
V02	*touching*	19.8–20.2	21.2–21.8	not judgeable	0.9 / 0.8	none
V03	*holding*	33.8–34.8	38.6–39.4	1.2 / 0.7	1.4 / 0.1	0.6 / 0.3
V04	*holding*	6.0–6.2	8.6–9.6	0.2 / 0.4	0.2 / 0.4	none
V04	*touching*	6.0–6.2	8.6–9.6	0.2 / 0.4	0.2 / 0.4	none
V04	*holding*	10.4–10.8	13.2–13.8	0.1 / 0.0	0.2 / 0.0	none
V04	*touching*	10.4–10.8	13.2–13.8	0.1 / 0.0	0.2 / 0.0	none
V05	*touching*	39.4–40.4	43.4–44.0	0.6 / 0.2	0.6 / 0.4	0.9 / 0.5
V05	*touching*	47.4–47.8	49.0–49.6	0.2 / 0.2	0.1 / 0.2	other predicate
V06	*touching*	23.2–23.6	25.2–26.2	not judgeable	not judgeable	not judgeable
V07	*touching*	29.2–29.8	34.8–35.4	0.4 / 3.0	other predicate	0.2 / 0.1
V08	*touching*	9.2–10.0	13.6–14.4	0.1 / 0.3	0.4 / 0.1	0.1 / 0.3

Use in this paper. The audit is descriptive: boundaries are not averaged into a reference, and no annotator is preferred. Timing scores use the original automatic labels throughout, and Section 3.4 reports them on the subset of events that both annotators record. Rescoring the fixed predictions against each annotator's records, over stated common support, is a possible later analysis.

E EARLIER INPUT AND INTERFACE CONTROLS

The input and interface controls preceded the causal remaining-time experiment and provide the contrasts in Section 3.6. They use the same eligible eight-video support, predict elapsed and remaining distances with a bidirectional head, and average onset- and end-interval errors. Figure 12 visualizes these comparisons. A later geometry and Ridge analysis examines the causal study's frozen POST interface. These analyses characterize access and readout choices on the same data as the causal experiment.

Input controls. PRE is the 128-dimensional paired readout before the source framewise block. Each input condition in Table 8 receives its own matched probe training runs, using three folds, three seeds, and 40 epochs. The bidirectional head has 42,370 parameters and outputs two distances. STATIC replaces each valid window's learned features by their mean, repeated across time; this is an offline train-and-test transformation, not test-only perturbation or a causal input. RAW pools actor and target features by mask area, concatenates them to 512 dimensions, and uses training-fold-only PCA to 128 dimensions. PCA is fitted on unique video/pair/frame entries with both participants visible. No explicit visibility indicator is appended. This raw path differs from the learned equal-occupied-patch query path, so its score does not isolate information lost during recognition training.

RAW outperforms learned PRE by 0.0591 seconds, with a paired interval $[-0.1056, -0.0209]$ and an improvement on seven of eight videos, but shows no reliable advantage over zero visual input or the label-only prior. The RAW–PRE difference points to an input-path effect whose source (pooling, training, or another mechanism) is open.

Interface controls. POST applies the source block in its original diagonal-attention mode, with sinusoidal positions reset every 16 native frames. PRE caches were stored in float16 and promoted to float32; the original source used uniformly sampled relation clips rather than dense native-frame blocks. The post-block tap therefore preserves a particular computation but not the original input distribution. A zero-PRE-through-block condition tests the contribution of source biases and position encoding without visual evidence; ZERO instead zeros the input to the temporal head.

The learned-minus-random difference is $+0.0074$ seconds at PRE and -0.0861 at POST. The difference between these contrasts is -0.0935 seconds, with a paired video-bootstrap interval $[-0.1856, -0.0114]$. In this exploratory analysis, the paired shift shows that the contrast depends on the interface, while the learned-minus-random contrast at POST remains uncertain, with interval

Table 8: **Bidirectional input controls favor the raw feature path over learned PRE.** Mean onset/end interval error (s), with three folds × three probe seeds per fitted condition (45 runs), on the same eight videos. The two paths also differ in pooling, so the contrast mixes pooling and recognition training.

Input	Error
Learned PRE	0.8307
Random PRE	0.8233
Zero visual	0.8539
Repeated static learned PRE	0.8880
Raw pooling + train-fold PCA	0.7716
Label-only prior	0.7840

Table 9: **The learned-minus-random contrast depends on the feature interface.** Mean onset/end interval error (s) from bidirectional heads; six conditions × three folds × three seeds (54 runs). PRE precedes the source framewise block and POST follows it.

Feature interface	Error
Learned PRE	0.8307
Random PRE	0.8233
Learned POST	0.7968
Random POST	0.8829
Zero PRE passed through learned block	0.8516
Zero input to temporal head	0.8539
Label-only prior	0.7840

$[-0.2211, +0.0383]$. Synthetic source-forward reconstruction matches logits exactly for tested 16- and seven-frame inputs, which verifies the implementation.

Fit and reproducibility diagnostics. An audit of the existing epoch-40 input-control checkpoints gives run-averaged training interval errors of 0.0231 (learned PRE), 0.0321 (random PRE), 0.0115 (RAW), 0.0342 (STATIC), and 0.4553 seconds (ZERO). These use run-level training aggregation rather than the eight-video test aggregation, so they are not directly comparable with the test means. Several visual heads therefore fit their training constraints. The final causal workflow passes 87 synthetic tests; all 27 trained heads also pass future-input invariance checks.

Representation geometry and a linear probe. The POST interface of the causal study was replayed on CPU (about one second, no training, no API calls) for all valid frames of the 22 fixed 32-frame windows that contain primary support frames: 691 context rows for 8 videos and 15 events, comprising 212 support contexts with remaining-time brackets and 479 unscored contexts whose status is unknown. The POST feature does not receive the predicate, so a frame queried under two predicates is one physical feature; the 691 rows correspond to 499 unique features. Figure 13 fits separate two-component PCAs on the 499 unique features of P and R, then colors support contexts by the bracket midpoint. This exploratory projection is distinct from the preprocessing used in cross-validation. Table 10 reports a ridge regression from a single frozen frame to the bracket midpoint with the original three video-disjoint folds as outer folds, an inner three-way video split for the penalty ($\lambda \in \{0.01, 0.1, 1, 10, 100, 1000\}$), and PCA, centering, and standardization fitted on the training split only; R^2 is reported pooled over frames and with events and videos weighted equally. The target is the remaining-time bracket midpoint; it differs from the interval-distance loss used by the causal head. Predictions are not clipped. Pooled R^2 uses the global mean of the pooled held-out targets in its denominator; the weighted variant uses their weighted global mean. Negative values therefore indicate error above that descriptive mean reference, not a comparison to a fitted training-fold-mean predictor.

The trained representation yields modest positive R^2 under both weightings, while the random-readout values are negative (Table 10). This describes the tested linear readout and sample, and the values depend on the aggregation. The support within-predicate/total variance ratios are 0.9557

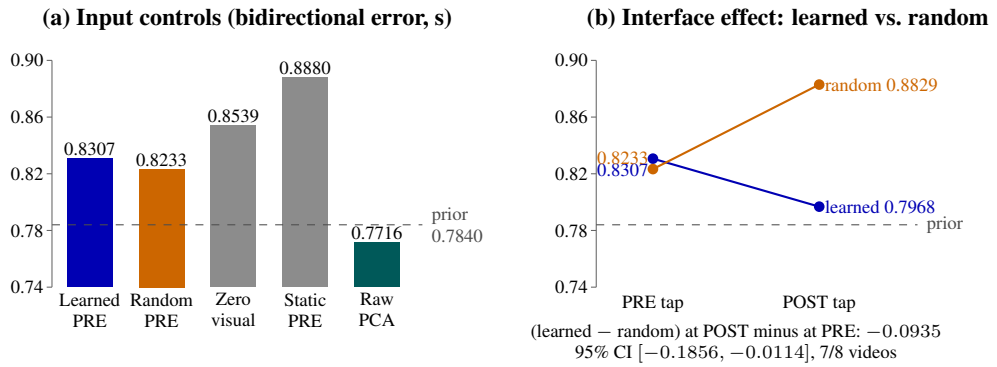

Figure 12: **Earlier bidirectional input and interface comparisons.** Values come from Tables 8 and 9. (a) Raw pooling with training-fold PCA has the lowest observed visual-input error; the label-only prior is shown dashed. (b) The learned-minus-random contrast changes between PRE and POST, with a paired interval for the shift excluding zero. Input paths differ, and the POST contrast alone remains uncertain. These exploratory results share the eight-video sample and predict both elapsed and remaining time.

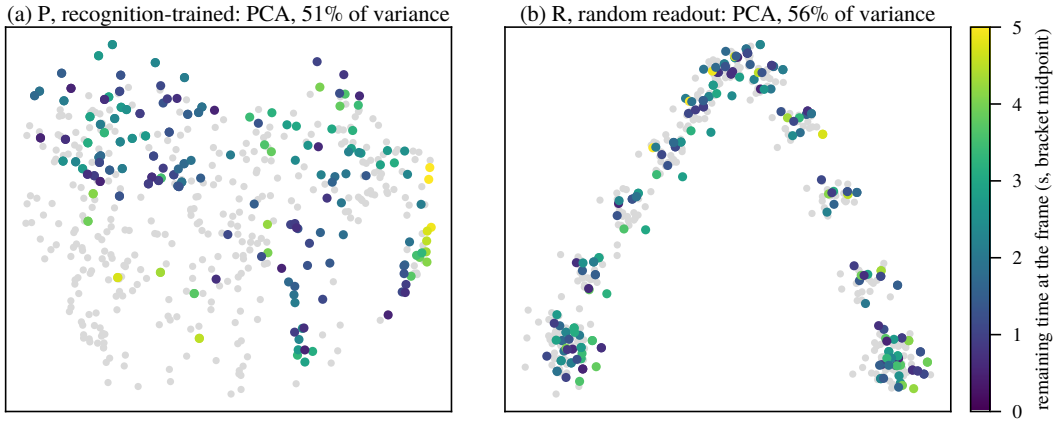

Figure 13: **Exploratory geometry of frozen POST features.** Separate PCAs use 499 unique visual features from 22 windows containing support. The 212 support contexts are colored by remaining-time bracket midpoint; grey marks provide the full feature context, including unscored observations. In R, the first component correlates with within-block position ($r = 0.98$); in P, 58% of the two-component variance is between videos. These associations describe the projection; Table 10 evaluates a linear readout.

for P and 0.9875 for R, computed over two overlapping predicates, not the 59-class source evaluation. No source-validation geometry export was available. These statistics and the two-dimensional projection describe the stated frozen interface only.

Table 10: **Remaining-time midpoint prediction with a linear readout.** Nested video-group Ridge regression from individual frozen POST features on 212 support contexts. R^2 is pooled or weighted equally by video and then event; MAE is in seconds. Negative R^2 means an error above that of the pooled target-mean reference (weighted mean for weighted R^2).

Input	R^2 pooled	R^2 weighted	MAE pooled (s)	MAE weighted (s)
P, top-10 PCs	$+0.053$	$+0.169$	0.897	0.802
P, all 128 dims	$+0.006$	$+0.142$	0.905	0.812
R, top-10 PCs	-0.025	-0.056	0.946	0.973
R, all 128 dims	-0.107	-0.106	0.962	0.972

F EVIDENCE SCOPE AND CLAIM BOUNDARIES

Table 11 distinguishes the learning targets and evidence discussed in Section 4. Table 12 maps the principal claims to their supporting evidence and limits. The pilot evaluates recognition, one temporal readout, a same-prefix replay, and a comparison of the automatic labels with two human annotators.

Table 11: **Learning targets and their evidence.** These emphases can coexist in the same object-based or foundation model. Relational dynamics models already pursue aspects of interaction state. Matched objective comparisons and broader transfer tests remain proposed research.

	Identity/localization targets	Interaction recognition	Interaction-state learning
Target emphasis	Entity identity, location, and persistence	The specified semantic or relational output	Relational state, relevant history, and possible changes
Example evidence	Localization and association performance	Correct labels, roles, or temporally localized relations	Event distinctions, prediction, and calibrated uncertainty
Further question	Which relational distinctions are accessible?	Which same-category event differences are accessible?	Which information supports additional tasks and contexts?
Comparison needed	Preserve object competence when changing objectives	Measure recognition alongside event judgments	Match backbone, data, observations, capacity, and adaptation
Scope of claim	The tested object tasks	The tested recognition tasks	Interaction-state access and transfer within stated uses; physical utility tested separately

Scope. The evaluation supplies the participants and assumes that the event is active. Participant discovery, event detection, state-transition prediction, interaction composition, intent, policy success, and safe deployment are outside its scope.

Table 12: **Claim–evidence map.** Principal claims, their supporting observations, and limits on interpretation. Numerical comparisons retain the reported protocols and aggregation.

Claim (where)	Evidence	Does not show
Object-task or category success can leave an event distinction open (§2.2); *logical*	counterexample: targets equal while a needed distinction varies	whether a given trained network keeps the distinction, which is checked per task
Participant conditioning improves source recognition (Table 1a, App. A); *measured*	$+19.39$ Recall@3 points over the actor-only query, paired video-bootstrap interval $[14.29, 24.95]$, three source seeds; late-fusion contrast crosses zero; ROI contrast $+3.08$ $[0.62, 5.63]$	that joint attention is the best mechanism; an effect of scene context; automatic participant discovery; broader transfer
The tested interface shows no reliable temporal gain (Table 1b, App. B); *pilot finding*	the three paired video-bootstrap intervals cross zero (App. B); P improves on R in 3/8 videos and on Z in 4/8; omitting single videos changes the sign of the R and Z point contrasts	equivalence; absence of temporal information; information removed by recognition training; deletion ranges as confidence intervals
The label-only prior has the lowest sample error (Table 1b)	0.6117 s against 0.7580–0.8106 s for the three probe conditions on the fixed sample	a population-level ordering: the P-minus-prior interval is $[-0.0201, +0.4905]$
The P–prior ordering survives the endpoint audit (Table 4)	deterministic range $[+0.0168, +0.3426]$ s over all shared integer endpoints inside the annotated brackets, with predictions and sample fixed	a confidence interval; label accuracy; endpoints outside the brackets; generalization to other samples
The recorded category is often retrievable from the temporal head's prefixes (Table 5, App. C); *measured, uncertain margin*	video-balanced Recall@3 44.80% $[23.35, 63.56]$ over the 212 contexts against a 25.00% frequency prior; paired difference $+19.80$ pp $[-4.63, +45.40]$	observable-state understanding; a history-length effect (sampling, positions, and task privileges differ from the full-event ranks); a reliable advantage over the prior
Input and interface choices affect temporal readout error (App. E, Fig. 12)	RAW $-$ learned PRE $= -0.0591$ $[-0.1056, -0.0209]$, 7/8 videos; post-minus-pre change of the learned-minus-random difference $= -0.0935$ $[-0.1856, -0.0114]$	information lost during recognition training; the RAW and learned query paths differ
Linear readout scores differ across conditions and aggregation (Table 10)	video-disjoint ridge R^2 0.05 pooled / 0.17 weighted for P against -0.02 / -0.06 for R; MAE 0.80–0.90 s for P	a reliable population-level advantage; no information in R; all information available in either representation
The recognition-rank association is imprecisely estimated (Fig. 6)	Spearman $\rho = 0.10$ over 15 events, 8-video block-bootstrap interval $[-0.47, 0.66]$; P ranks the recorded predicate in its top three for 9/15 events	independence or absence of association; the interval admits substantial correlations in either direction
The automatic labels are close to human ones where both annotators record an event (Table 2, App. D); *descriptive*	both final human versions record 9 of the 15 scored events; median start and end differences of 0.2 s from each annotator, against 0.1 and 0.0 s between annotators; six events not confirmed by both	label accuracy for the six unconfirmed events; a consensus reference; a human-label model score
Explicit relational targets improve transfer or cost (§2.2, §5); *hypothesis*	cognitive motivation and relational-modeling precedents; no SAM4D objective comparison	superiority; general pretraining could match or exceed it under matched conditions

G VISUAL COVERAGE OF THE EIGHT-VIDEO PILOT

Figures 14 and 15 show every evaluation video, all nine participant pairs, and all 15 scored events with full RGB frames, so that the two cases selected for Figure 7 can be read against the complete support. Frame selection follows a fixed rule based on scored support and endpoint ranges, independent of prediction quality. The displayed names of two targets (a cat tree and a globe) are display labels for objects whose recorded data category is *toy*. Curves and errors are the pilot results on the automatic labels.



Figure 14: **Complete eight-video atlas (1/2).** V01–V04 show full RGB frames with supplied participant contours (actor blue, target orange) and each scored event. Each chart reports the complete-event recognition rank of the annotated predicate under P, and temporal errors for P/R/Z/prior in seconds. Curves average predictions across seeds; errors average per-frame errors within each seed, then seeds. Gray bands are the original automatic endpoint targets. Curves stop outside exported scored frames and break at 32-frame input-window resets. Frame selection follows a fixed rule based on scored support and endpoint ranges, independent of prediction quality.

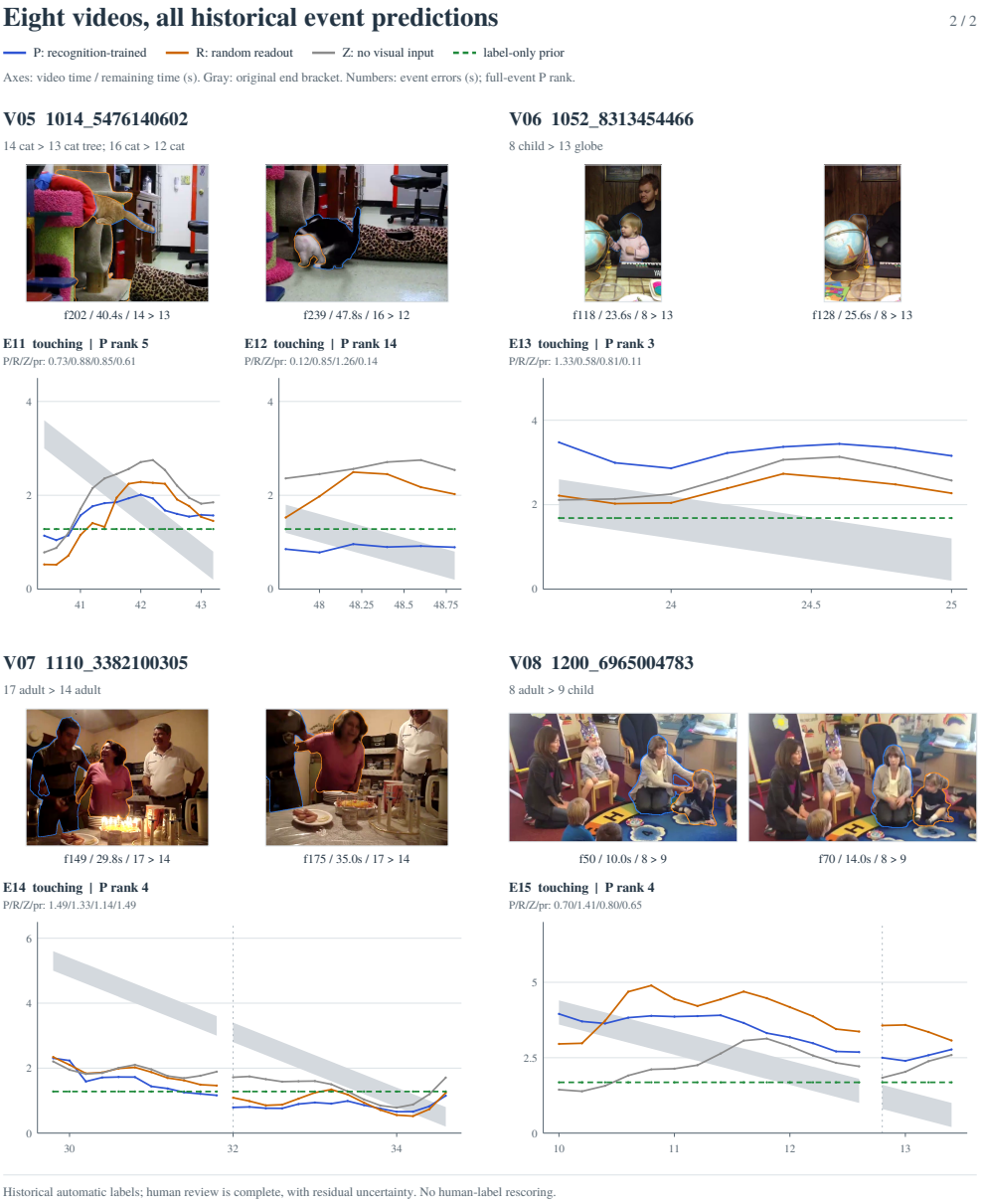

Figure 15: **Complete eight-video atlas (2/2).** V05–V08 complete coverage of all eight videos, nine participant pairs, and fifteen original events. V05 contains two pairs, shown separately by the frame labels and event identifiers. Encoding follows Figure 14; each video's charts share a vertical scale. The standalone eight-page case atlas provides larger frame strips, whole-video support timelines, seed ranges, and the separately recorded human-review status. Unknown or unsupported frames have no extrapolated predictions.