WHAT HAPPENS NEXT? FUTURE-BASED STREAMING HUMAN–OBJECT INTERACTION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Many video human–object interaction are defined by the changes they induce rather than by their initial appearance: a person beginning to push a box can look identical to one beginning to pull it until the box moves. Offline detectors can wait for this consequence, but systems that interpret interactions online cannot. We study streaming human–object interaction detection under a bounded-lag protocol. A predicate is emitted from the current prefix and may be revised once after a fixed lag, while tubes, identities, and temporal boundaries remain frozen. All methods therefore receive the same proposals and later frames, isolating predicate interpretation from delayed localization. We show that an unconditioned rollout of a frozen latent video model produces a shared average future and cannot distinguish candidate predicates. Conditioning the model on each candidate yields distinct future distributions, but only over a finite horizon in which the interaction remains identifiable. RIVET exploits this signal with predicate-conditioned simulators, retrieval supervision for unrealized branches, and continuation matching against the future that actually occurs. A separation gate abstains when candidates are indistinguishable, while learn-then-test calibration controls the expected degradation of revision relative to a delayed reference. Under full-video evaluation, RIVET achieves relation-detection mAP of 12.22% on VidOR, compared with 11.98% for the action-expert baselines.

1 INTRODUCTION

Video human-object interaction detection localizes people and objects, identifies the predicates between them, and maintains these associations over time (Shang et al., 2019; Chiou et al., 2021). It extends image-level HOI detection (Lei et al., 2024; Geng et al., 2025; Hu et al., 2025) to the temporal domain and is closely related to video visual relation detection and dynamic scene graph generation (Nag et al., 2023; Wang et al., 2024; Wei et al., 2024; Jiang et al., 2024; Li et al., 2025b; Wang et al., 2025b; Le et al., 2026). Systems that must react while an interaction unfolds depend on this capability, including assistive robots (Koppula & Saxena, 2016; Bharadhwaj et al., 2024), augmented-reality agents (Grauman et al., 2022; 2024), and video monitoring (Wu et al., 2024b). In such settings, recognizing an interaction only after its full consequence becomes visible may be too late. A useful system infers the interaction from partial observations (Kong & Fu, 2022), maintains stable human-object associations, and updates its prediction as new frames arrive.

This motivates Streaming Human-object Interaction (SHOI) detection (Fig. 1(b)), where a system emits predictions online under a bounded observation delay. Streaming HOI differs from offline HOI detection (Fig. 1(a)), which may inspect the complete interaction and use future frames around the target interval. Online action detection, streaming perception, and streaming video language models (De Geest et al., 2016; Li et al., 2020; Chen et al., 2024; Qian et al., 2025; Wang et al., 2025a) impose similar latency constraints, and recent online detectors even admit a small future latency (Pang et al., 2025). These methods label frame-level actions, localize objects, or generate language responses, and they do not revise the predicate of a tracked human-object pair once later frames arrive. Streaming HOI is also more than applying an offline detector to a shorter clip, since the system must determine what can be inferred from the current prefix, when additional evidence should be used, and whether an earlier prediction should be revised. To our knowledge, existing HOI benchmarks and video relation detection methods provide no unified protocol for bounded-lag predicate revision.

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.

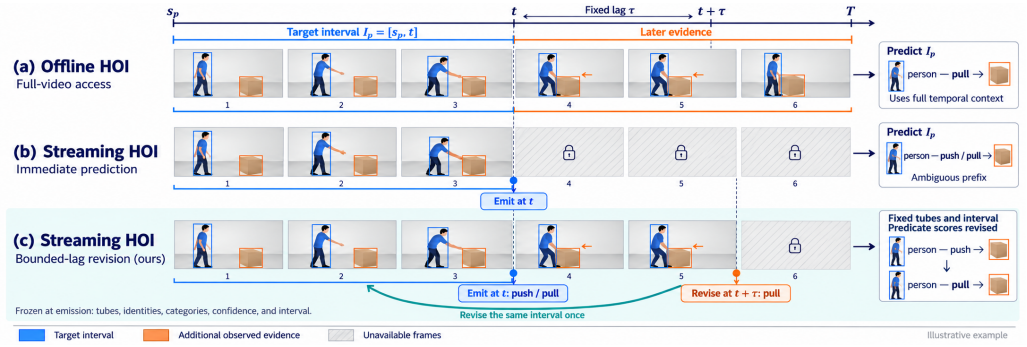

Figure 1: Offline and streaming HOI detection. (a) Offline HOI uses full-video context. (b) Streaming HOI predicts from the available prefix, which may leave predicates ambiguous. (c) Our bounded-lag protocol emits a prediction at t and permits one revision at $t + \tau$ for the interval $I_p = [s_p, t]$.

The problem is hardest for predicates defined by changes in motion or object state rather than by initial appearance (Souček et al., 2022; Xue et al., 2024). A person beginning to push a box looks similar to one beginning to pull it until the box responds. Putting and taking, or opening and closing, are similarly ambiguous from a short prefix. Fine-grained temporal benchmarks show that such distinctions depend on temporal evidence rather than static appearance (Goyal et al., 2017; Sevilla-Lara et al., 2021; Cai et al., 2024). The evidence that distinguishes such predicates may arrive only after the prediction has been emitted, even when the interacting entities are correctly localized.

We formulate this setting as bounded-lag predicate revision. A system emits a prediction from the history available at the current time and may revise the predicate scores once after a fixed lag, using the continuation that has arrived by then. The human and object tubes, identities, categories, proposal confidence, and temporal boundaries stay frozen at emission time. This isolates delayed predicate interpretation from delayed localization, since any improvement after revision must come from reinterpreting the same human-object pair, as shown in Fig. 1(c).

Our hypothesis is that pretrained world models (Bardes et al., 2024; Hafner et al., 2025; Bar et al., 2025; Zhou et al., 2025; Assran et al., 2025) provide a useful inductive bias. A discriminative recognizer must learn from labeled examples how a continuation supports each predicate, whereas a pretrained latent video predictor has additionally learned how visual scenes evolve (Wu et al., 2024a; NVIDIA et al., 2025). Conditioning this transition structure on the interacting pair and a candidate predicate yields competing hypotheses about future evidence. Matching these hypotheses against the continuation that actually arrives may improve predicate estimation at shorter delays and with fewer labels. The idea relates to recognition through generation (Li et al., 2023; 2025a), action and scene graph anticipation (Girdhar & Grauman, 2021; Zhao et al., 2024; Peddi et al., 2024; 2025), and latent-action modeling (Schmidt & Jiang, 2024; Bruce et al., 2024; Ye et al., 2025). Anticipation forecasts future actions or relations, whereas we use candidate-specific future distributions to reinterpret an interval that has already been emitted.

We introduce RIVET, a future-conditioned revision method for streaming HOI. At emission time it selects candidate predicates and uses a frozen latent video predictor (Oquab et al., 2024; Zhou et al., 2025) with pair- and predicate-specific conditioning to produce a future evidence distribution for each candidate. The observed continuation supervises the realized predicate, while retrieved transitions from comparable training scenes regularize unrealized ones. When the continuation arrives, RIVET compares the observed pair-level evidence with the candidate distributions in a fixed evidence space and applies a residual correction to the delayed reference. These rollouts and retrieved targets are observational predictive proxies and do not identify causal interventions (Pearl, 2009).

RIVET also estimates candidate separation and suppresses revision when the predicted futures are insufficiently distinguishable or the continuation cannot be reliably associated with the tracked pair, in the spirit of selective prediction (Geifman & El-Yaniv, 2017). Separation alone does not guarantee a correct revision, so learn-then-test calibration (Angelopoulos et al., 2025), a member of the distribution-free risk-control family (Bates et al., 2021; Angelopoulos et al., 2024), selects the revision strength and gating thresholds on held-out videos. Under the stated exchangeability assumptions, calibration controls the expected fraction of corrupted reference predictions and the

expected per-video average-precision degradation. If no nontrivial configuration satisfies the tolerances, RIVET returns the delayed reference unchanged.

We evaluate on VidOR (Shang et al., 2019), VidVRD (Shang et al., 2017), and Action Genome (Ji et al., 2020). We measure the predicate bottleneck, performance across observation delays, label efficiency, and matched-lag improvements over discriminative and retrieval-based baselines. Unconditioned rollouts, random predictors, shuffled labels, and alternative continuation scorers test whether the gains arise from pretrained candidate-conditioned dynamics. We also analyze separation, calibration, computation, and failure cases.

Contributions. (i) We formulate bounded-lag streaming HOI, in which predicates may be revised once after a fixed delay. (ii) We devise a matched-lag protocol that gives all methods the same continuation, separating predictive modeling from the benefit of seeing more frames. (iii) We present **RIVET**, which matches pair- and predicate-conditioned future evidence distributions against the observed continuation and uses separation-aware calibration.

2 PRELIMINARIES

Problem definition. Let $V = (x_1, x_2, \ldots)$ denote a video stream, and let $\mathcal{H}_t$ contain the frames and tracking states available up to time t. A grounding module emits interaction proposals. Each proposal p consists of human and object tracklets $B^h_{1:t,p}$ and $B^o_{1:t,p}$, an object category o_p, a confidence κ_p, and a fixed-length interval $I_p = [s_p, t]$ emitted at a fixed stride. Its output is $(B^h_{I_p,p}, B^o_{I_p,p}, o_p, I_p, \boldsymbol{\sigma}_p)$, where $\boldsymbol{\sigma}_p$ contains scores over the interaction vocabulary $\mathcal{A}$.

Bounded-lag protocol. Some interactions have similar initial appearances but become distinguishable only through their subsequent effects, such as the object motion that separates pushing from pulling. A streaming detector must therefore predict before the interaction is complete and may need to revise that prediction when later evidence arrives. We model this setting as follows. The initial prediction uses only $\mathcal{H}_t$; after a fixed lag τ, one revision may use $\mathcal{H}_{t+\tau}$. Boxes, identities, category, confidence, and interval remain fixed, so tracking provides evidence about the same pair. All methods use the same proposals, matching procedure, and duplicate handling. Revised scores label I_p, even if the interaction changes during the continuation. We evaluate emission and revision separately and retain the emitted prediction when no continuation is available.

Labels and reference predictions. Let $Y_p \subseteq \mathcal{A}$ contain the annotated interactions for proposal p. Fully annotated background proposals receive a null label, whereas unannotated or ignored proposals are excluded from supervision. We draw A_p uniformly from Y_p, allowing multiple physical relations to co-occur. Training averages over all annotated labels without treating a co-occurring positive as a negative, and inference returns a ranked list per pair.

The prefix recognizer estimates $\pi^0_p(a) \approx \Pr(A_p = a \mid \mathcal{H}_t, I_p)$ from the prefix alone. The delayed reference $b_p(a) = b(a \mid \mathcal{H}_{t+\tau}, I_p)$ labels the same interval using the extended window and the same proposals, tracks, and encoder as the evaluated method. At zero lag or when the continuation is unavailable, $b_p = \pi^0_p$.

Latent dynamics and pair-specific evidence. We use a DINO-WM model with encoder E and predictor P. A context clip ending at t is encoded as $z_t = E(c_t) \in \mathbb{R}^{N \times d}$. Target positions specify the future representation at lag τ, and the corresponding observed clip ends no later than $t + \tau$ and is encoded as $z_{t+\tau}$ using the same temporal sampling. The base weights remain frozen, and P is adapted only through additional trainable parameters.

Comparing full $N \times d$ latents is costly and does not isolate the interacting pair. We therefore cache a query Q_p from $\mathcal{H}_t$ that represents the human, object, their union, and their observed geometry and motion, and define pair-level evidence $u^s_p \in \mathbb{R}^{d_u}$ by

$$u^s_p = R_\xi(z_s; Q_p) = \mathrm{Pool}(\mathrm{CrossAttn}_\xi(Q_p, z_s, z_s)), \qquad s \in \{t, t+\tau\}. \tag{1}$$

Predicted and observed futures use the same query, which contains neither candidate labels nor future boxes. Tracking verifies that the pair remains observable without modifying Q_p. After prefix training, we freeze R_ξ and standardize its outputs using training statistics with a positive scale floor, keeping the evidence targets fixed during later stages.

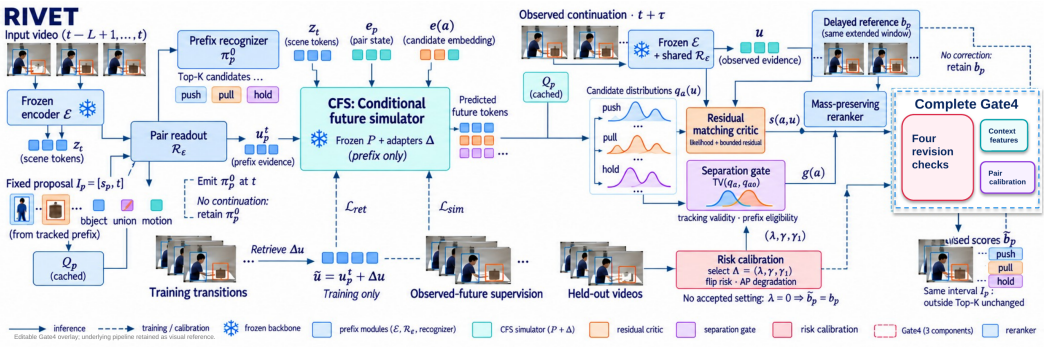

Figure 2: Overview of RIVET. (a) The video prefix yields pair-specific features and top-K interaction candidates. (b) A predicate-conditioned simulator predicts future evidence distributions for each candidate. (c) When the continuation arrives, a shared critic matches observed and predicted evidence, while a separation gate controls corrections to the delayed reference.

Predicate-conditioned future evidence. Similar prefixes may correspond to different interactions, such as pushing and pulling, whose differences emerge through subsequent object motion. We therefore model the future evidence for each candidate predicate and compare it with the observed continuation. This transfers temporal knowledge from a pretrained video prediction model to interaction recognition when labeled continuations are scarce.

Let $h_p \subseteq \mathcal{H}_t$ denote the prefix information and $u = u_p^{t+\tau}$ the continuation evidence. With true label prior $p_0(a \mid h_p)$ and conditional evidence distribution $p_\star(u \mid h_p, a)$, Bayes' rule gives

$$p_\star(a \mid h_p, u) = \frac{p_0(a \mid h_p)\, p_\star(u \mid h_p, a)}{\sum_{a' \in \mathcal{A}} p_0(a' \mid h_p)\, p_\star(u \mid h_p, a')}. \tag{2}$$

The likelihood favors predicates that better explain the continuation. If all conditional distributions coincide, the likelihood terms cancel and the posterior equals the prior. This motivates predicate-conditioned prediction, whose usefulness depends on predictive fidelity.

Proposition 1 (Separation depends on predictive fidelity)**.** *For a fixed h_p, let q_a and $q_{a'}$ approximate $p_{\star,a} = p_\star(\cdot \mid h_p, a)$ and $p_{\star,a'} = p_\star(\cdot \mid h_p, a')$, respectively, and define $\delta_\star = \mathrm{TV}(p_{\star,a}, p_{\star,a'})$. Then*

$$|\mathrm{TV}(q_a, q_{a'}) - \delta_\star| \le \mathrm{TV}(q_a, p_{\star,a}) + \mathrm{TV}(q_{a'}, p_{\star,a'}). \tag{3}$$

Large predicted separation may therefore reflect model error and does not identify which predicate explains the observed evidence. RIVET addresses this limitation by matching candidate futures to the continuation, using separation to gate corrections, and calibrating expected revision damage without assuming perfect simulation. Because the delayed reference b_p already observes the continuation, the correction is learned as a discriminative residual.

3 Method

We propose RIVET (Fig. 2), which evaluates interaction hypotheses through their predicted futures. The conditional future simulator learns pair- and predicate-specific future distributions from observed and retrieved transitions (§3.1). A shared critic matches these distributions to the observed continuation, while a separation gate controls residual corrections to the delayed reference (§3.2). Staged training and video-level risk calibration determine the correction strength and gating thresholds, controlling expected degradation during streaming inference (§3.3).

3.1 Predicate-Conditioned Future Simulation

CFS predicts future evidence for each candidate interaction, conditioned on both the target pair and the predicate. Observed continuations supervise annotated predicates, while retrieved transitions provide targets for alternative candidates.

Pair-conditioned prediction. A scene may contain multiple people and objects, so the predicate alone does not identify which interaction to predict. The prefix recognizer selects $K_p = \mathrm{TopK}(\pi_p^0)$.

For each $a \in K_p$, the predictor generates $\widetilde{z}_{p,\tau}(a) = P_\Delta([z_t; e_p; e(a)], \tau)$, where $e_p = F(Q_p, u_p^t)$ represents the pair and its observed state, $e(a)$ is the predicate embedding, and Δ denotes trainable low-rank adapters in the frozen predictor.

Even for a fixed pair and predicate, several outcomes are possible, such as weak or strong object displacement after a push. We therefore represent future evidence with a Gaussian mixture rather than a single mean. A head maps $R_\xi(\widetilde{z}_{p,\tau}(a); Q_p)$ to $q_\theta(u \mid h_p, a, \tau) = \sum_{m=1}^{M} w_{p,a,m}\mathcal{N}(u; \mu_{p,a,m}, \Sigma_{p,a,m})$, $\Sigma_{p,a,m} = \mathrm{diag}(\boldsymbol{v}_{p,a,m})$. Mixture weights use softmax, and each variance coordinate is parameterized as $v_{\min} + (v_{\max} - v_{\min})\,\mathrm{sigmoid}(r)$, with fixed $0 < v_{\min} < v_{\max}$, preventing singular components and unbounded component variance.

Observed-future supervision. We train the simulator by the negative log-likelihood of the observed continuation:

$$\mathcal{L}_{\mathrm{sim}}(p) = -\frac{1}{|Y_p|}\sum_{a \in Y_p} \log q_\theta(u_p^{t+\tau} \mid h_p, a, \tau). \tag{4}$$

Training includes annotated predicates outside K_p; inference uses only the selected candidates.

Retrieval-based supervision. Alternatives in $K_p \setminus Y_p$ have no direct target in the current proposal. We thus retrieve transitions with the corresponding label from other training videos, matching object category, horizon, and fixed appearance or motion descriptors. Retrieval excludes the query video, calibration data, and test data; a development-set distance cutoff rejects poor matches.

For a retrieved neighbor r with prefix endpoint t_r, we transfer its evidence change $\Delta u_r = u_r^{t_r+\tau} - u_r^{t_r}$ to the query as $\widetilde{u}_{p,r} = u_p^t + \Delta u_r$. Let $\mathcal{R}_p(a)$ contain the accepted neighbors, let $\omega_{p,r}^a$ be their normalized distance weights, and let $\mathcal{B}_p$ contain alternatives with nonempty neighbor sets. For $\mathcal{B}_p \neq \varnothing$, we use

$$\mathcal{L}_{\mathrm{ret}}(p) = -\frac{1}{|\mathcal{B}_p|}\sum_{a \in \mathcal{B}_p}\sum_{r \in \mathcal{R}_p(a)} \omega_{p,r}^a \log q_\theta(\widetilde{u}_{p,r} \mid h_p, a, \tau). \tag{5}$$

The loss is zero when $\mathcal{B}_p$ is empty. This transfer assumes comparable evidence changes in similar contexts and provides observational, not causal, supervision. We train the simulator by minimizing $\mathbb{E}_p[\mathcal{L}_{\mathrm{sim}} + \lambda_{\mathrm{ret}}\mathcal{L}_{\mathrm{ret}}]$ and then freeze it before training the reranker.

3.2 Separation-Gated Residual Correction

At revision time, the observed evidence $u = u_p^{t+\tau}$ is used by the delayed reference b_p. Multiplying b_p by a candidate likelihood would therefore count the same evidence twice. RIVET instead learns a residual compatibility score and applies it only when candidate futures are sufficiently distinct.

Residual matching. For each candidate, a shared critic adds a bounded compatibility correction to each mixture component: $k_{p,a,m}(u) = \log \mathcal{N}(u; \mu_{p,a,m}, \Sigma_{p,a,m}) + c\tanh\phi_\eta(u, \mu_{p,a,m}, \log \boldsymbol{v}_{p,a,m}, u_p^t)$. It then aggregates the components as $s_p(a,u) = \log\sum_{m=1}^{M} w_{p,a,m}\exp k_{p,a,m}(u)$. The critic shares parameters across candidates, receives no predicate embedding or label index, and is invariant to component order. The constant $c \geq 0$ is fixed on development data; $c = 0$ recovers mixture-likelihood matching.

Because b_p already observes u, we train the critic to correct the reference rather than to perform another Bayesian update. Let $Y_p^K = Y_p \cap K_p$ and $b_p^K(a) = b_p(a)/\sum_{a' \in K_p} b_p(a')$. For $Y_p^K \neq \varnothing$, we optimize only η using

$$\mathcal{L}_{\mathrm{crit}}(p) = -\tfrac{1}{|Y_p^K|}\sum\nolimits_{a \in Y_p^K} \log\left(\frac{b_p^K(a)\exp s_p(a,u)}{\sum_{a' \in K_p} b_p^K(a')\exp s_p(a',u)}\right). \tag{6}$$

This loss learns how the continuation reweights the delayed reference within the candidate set.

Separation-aware gating. A residual should not be applied when candidate futures are nearly indistinguishable. Let $a_0 = \arg\max_{a \in \mathcal{A}} b_p(a)$ be the reference leader. If $a_0 \notin K_p$ or $|K_p| < 2$, RIVET returns b_p. Otherwise, for $q_a = q_\theta(\cdot \mid h_p, a, \tau)$, we estimate the separation between q_a and q_{a_0} by sampling from $r_a = (q_a + q_{a_0})/2$: $\widehat{\delta}_p(a) = \frac{1}{S}\sum_{s=1}^{S} \frac{|q_a(U_s) - q_{a_0}(U_s)|}{q_a(U_s) + q_{a_0}(U_s)}$, $U_s \overset{\mathrm{iid}}{\sim} r_a$.This estimator is unbiased for $\mathrm{TV}(q_a, q_{a_0})$. Conditional on the fitted model, Hoeffding's inequality bounds its

sampling error by $\sqrt{\log(2/\zeta)/(2S)}$ with probability $1-\zeta$ per comparison; model misspecification is not covered by this bound.

Let v_p indicate valid continuation evidence and reliable tracking of both entities, and let χ_p be a prefix-only eligibility rule fixed before risk calibration. If $v_p = 0$, RIVET returns b_p without scoring. Otherwise,

$$g_p(a) = v_p \chi_p \mathbf{1}\{\widehat{\delta}_p(a) > \gamma\} \min\{\widehat{\delta}_p(a)/\gamma_1, 1\}, \qquad \gamma \geq 0, \quad \gamma_1 > 0. \tag{7}$$

Here γ is the minimum separation for revision and γ_1 controls gate saturation. Separation permits a correction but does not establish its correctness.

Mass-preserving revision. For $a \in K_p$, define the correction relative to the reference leader as

$$d_p(a) = \lambda g_p(a)\big[s_p(a,u) - s_p(a_0,u)\big]\mathbf{1}\{a \neq a_0\}, \qquad \lambda \geq 0. \tag{8}$$

Let $m_p = \sum_{a \in K_p} b_p(a)$ be the reference probability mass inside K_p. We redistribute only this mass:

$$\widetilde{b}_p(a) = \frac{m_p\, b_p^K(a) \exp d_p(a)}{\sum_{a' \in K_p} b_p^K(a') \exp d_p(a')} \quad (a \in K_p), \qquad \widetilde{b}_p(a) = b_p(a) \quad (a \notin K_p). \tag{9}$$

Scores outside K_p remain unchanged, and a closed gate preserves the candidate's odds relative to a_0. If all gates are closed or $\lambda = 0$, the update exactly recovers b_p.

Gate4 scoring. We use Gate4 as a unified module combining four revision checks with context-aware pair calibration. It applies a shared pair-level logit offset to the relation scores using candidate statistics, detector confidences, interval duration, and same-predicate context features. Upstream ablations retain the complete Gate4 module.

3.3 Training, Calibration, and Streaming Inference

Training. We first train the recognizers and shared prediction head and freeze them. We then train CFS with Eqs. 4 and 5, freeze the simulator, and train the critic with Eq. 6.

Risk calibration. Predicted separation indicates when candidate futures differ, but it does not guarantee that revision is correct. We therefore select $\Lambda = (\lambda, \gamma, \gamma_1)$ on independent calibration videos. For a video V, let $\mathcal{Q}_V$ be its evaluable proposals, $a_p^b = \arg\max_{a \in \mathcal{A}} b_p(a)$, and $a_p^\Lambda = \arg\max_{a \in \mathcal{A}} \widetilde{b}_p(a)$. We control

$$L_{\text{flip}}(V;\Lambda) = \frac{\sum_{p \in \mathcal{Q}_V} \mathbf{1}\{a_p^b \in Y_p,\ a_p^\Lambda \notin Y_p\}}{\max\{1, |\mathcal{Q}_V|\}}, \qquad L_{\text{AP}}(V;\Lambda) = \Big[\text{AP}_V(b) - \text{AP}_V(\widetilde{b})\Big]_+. \tag{10}$$

The first risk measures correct-to-incorrect flips; the second measures the decrease in per-video AP. Both lie in $[0,1]$ and use the same detections, matching, and class averaging for both predictors.

We fix J candidate configurations using development data. On n independent calibration videos, let $\widehat{R}_j(\Lambda)$ be the empirical risk for $j \in \{\text{flip}, \text{AP}\}$. We accept only configurations satisfying

$$U_j(\Lambda) = \min\left\{1, \widehat{R}_j(\Lambda) + \sqrt{\frac{\log(2J/\beta)}{2n}}\right\}, qquad \mathcal{C} = \{\Lambda : U_j(\Lambda) \leq \alpha_j \text{ for both } j\}. \tag{11}$$

Among accepted configurations, we choose the one with the largest empirical AP improvement, breaking ties by smaller λ. If none is accepted, $\lambda = 0$ recovers b_p.

Conditioned on the fitted models, preprocessing rules, and candidate grid, if calibration videos are i.i.d. with test videos, then for $R_j(\Lambda) = \mathbb{E}_V[L_j(V;\Lambda)]$,

$$\Pr_{\mathcal{D}_{\text{cal}}}\Big\{R_{\text{flip}}(\widehat{\Lambda}) \leq \alpha_{\text{flip}}, \quad R_{\text{AP}}(\widehat{\Lambda}) \leq \alpha_{\text{AP}}\Big\} \geq 1 - \beta. \tag{12}$$

This guarantee controls expected degradation over the video population, not individual videos or dataset-level mAP. Multiple lags share the same multiplicity budget.

Streaming inference. At time t, RIVET computes prefix scores, caches the proposal and query, and generates candidate mixtures from $\mathcal{H}_t$. At time $t+\tau$, it computes the delayed reference and continuation evidence, checks validity and leader coverage, and applies Eq. 9 once. If the continuation is unavailable, it retains the prefix prediction.

4 EXPERIMENTS

Table 1: Main results (%). Published rows keep their original tracks and coverage; local rows share the action expert's outputs. (a) Full-video relation detection, with validation calibration on VidOR and local detector tracks on VidVRD. (b) AG PredCls without graph constraint on 51,052 eligible frames, not rank-comparable with published results. (c) Matched observation budget, $\tau = 0.5$ s. (d) Independent prefix and delayed reference heads on VidOR with shared proposals and fusion.

(a) Full-video relation detection

	VidOR			VidVRD		
Method	mAP	R@50	R@100	mAP	R@50	R@100
Published reference results						
HCM-RG (Wei et al., 2024)	10.44	9.74	11.23	29.68	17.97	21.45
VrdONE (Jiang et al., 2024)	11.86	11.13	14.21	31.33	18.20	21.61
VrdONE-X (+CLIP) (Jiang et al., 2024)	12.17	11.41	14.55	–	–	–
Controlled local comparisons						
Action expert	11.98	11.50	14.56	24.65	15.12	17.75
RIVET + Gate4 (native scores)	12.20	11.42	14.50	25.64	14.73	17.89
RIVET + Gate4 (conditional fusion)	12.22	11.44	14.50	26.01	15.23	18.02

(b) Action Genome PredCls, no graph constraint

Method	R@20	R@50	R@100	mR@20	mR@50	mR@100
Published standard-protocol results						
ARN (Wang et al., 2025b)	96.7	99.9	–	–	–	–
VISA (Li et al., 2025b)	–	–	–	89.1	99.8	–
UNO (ViT-L/14) (Le et al., 2026)	98.1	100.0	–	–	–	–
Our eligible-frame evaluation (51,052 frames)						
Action expert	90.78	98.65	99.88	69.09	93.49	99.66
RIVET + Gate4	90.78	98.66	99.88	69.09	93.49	99.66

(c) Matched observation budget, $\tau = 0.5\,s$

Method	Window	VidOR mAP	VidVRD mAP	AG mR@20
Prefix recognizer	$\mathcal{H}_t$	3.07	3.32	–
STTran	$\mathcal{H}_{t+\tau}$	–	–	48.09
TEMPURA	$\mathcal{H}_{t+\tau}$	–	–	64.77
Qwen2.5-VL-7B	$\mathcal{H}_{t+\tau}$	0.00	–	–
Delayed reference, frozen encoder	$\mathcal{H}_{t+\tau}$	5.45	12.18	–

(d) VidOR, independent reference heads

Method	mAP	R@50	R@100	CD-15 AP
Prefix-only head	5.31	5.83	9.42	3.14
Real-future delayed head	5.45	6.03	9.42	3.63

4.1 EXPERIMENTAL SETUP

Datasets and metrics. We evaluate VidOR (Shang et al., 2019) (80 object categories, 50 predicates) on its 835 validation videos, since test annotations are not public, and VidVRD (Shang et al., 2017) (132 predicates) on its 200 test videos. Both use relation-detection mAP and R@50/100. On Action Genome (AG) (Ji et al., 2020), which annotates 26 attention, spatial, and contact predicates, we evaluate PredCls without a graph constraint on 51,052 eligible frames from 1,773 videos. For AG we report R@K and mR@K with $K \in \{20, 50, 100\}$. This subset departs from the standard protocol, so published results are listed separately. All metrics are percentages. Consequence-defined subsets and uncertainty estimates appear in Appendix A.1.

Baselines and protocol. RIVET revises the scores of a discriminative action expert and shares its proposals, tracks, matching, and duplicate handling, so any difference arises from predicate scoring alone. Full-video evaluation follows the standard protocol and allows comparison with published

systems, but it does not test bounded lag. Matched-lag evaluation restricts every method to $\mathcal{H}_t$ or $\mathcal{H}_{t+\tau}$ with $\tau = 0.5$ s (Table 1(c)). Operating points are selected on VidOR validation videos and on a VidVRD development split. Published results for HCM-RG (Wei et al., 2024) and VrdONE (Jiang et al., 2024) on relation detection, and for ARN (Wang et al., 2025b), VISA (Li et al., 2025b), and UNO (Le et al., 2026) on AG, retain their original pipelines. They serve as context rather than controlled comparisons. For matched-lag evaluation we add STTran (Cong et al., 2021) and TEMPURA (Nag et al., 2023) on AG, the only dataset they support, and Qwen2.5-VL-7B (**?**) on VidOR. ST-HOI (Chiou et al., 2021) is omitted because its weights are no longer public.

Implementation details. The latent predictor is DINO-WM (Zhou et al., 2025) with a frozen DINOv2 encoder (Oquab et al., 2024). Its predictor is pretrained on VidOR, and its base weights remain frozen while rank-8 adapters are trained. On VidOR we use eight observed frames at 224×224 resolution, $K = 10$ candidates, $M = 2$ mixture components, and evidence dimension $d_u = 256$, with learning rate and weight decay both set to 10^{-4}. Gate4 denotes the complete calibration module of the action expert, which combines the four revision checks with context-aware pair calibration. It is kept intact in every ablation, and its scoring parameters are fitted separately for each dataset. On VidOR we apply conditional score fusion, in which candidates untouched by revision keep their native expert scores and revised candidates use detector-product scores before pair calibration.

4.2 Main Results

Full-video relation detection. Table 1(a) shows that RIVET with conditional fusion raises VidOR mAP from 11.98 to 12.22, and the native-score variant reaches 12.20. The paired-video bootstrap in Appendix A.1 gives a gain of +0.24 pp with a 95% interval of [+0.09, +0.42] for conditional fusion and +0.21 pp with an interval of [+0.07, +0.38] for native scores. These intervals exclude calibration refitting and training-seed variability. Recall does not improve, as R@50 falls from 11.50 to 11.44 and R@100 from 14.56 to 14.50, which is consistent with a reranking effect among relations that the expert already retrieves. The resulting VidOR mAP is on par with VrdONE-X, although that comparison is not controlled. On VidVRD, native-score revision improves mAP by 0.19 pp over the expert. The local VidVRD expert trails the published systems by more than 5 pp mAP, so this result measures revision over a weaker base detector. The gains do not concentrate on consequence-defined predicates. On the candidate CD-15 subset of VidOR, AP decreases by 0.25 pp and its interval includes zero (Appendix A.1).

Action Genome. Table 1(b) shows that RIVET leaves AG recall unchanged at two decimals, with no metric moving by more than 0.005 pp. The future branch therefore has no measurable effect on frame-level PredCls in our eligible subset. The published rows use the standard protocol and are not rank-comparable with our subset.

Matched observation budget. Table 1(c) isolates the value of the continuation. Extending the window from $\mathcal{H}_t$ to $\mathcal{H}_{t+\tau}$ raises mAP from 3.07 to 5.45 on VidOR and from 3.32 to 12.18 on VidVRD, which confirms that the continuation carries predicate information that the prefix lacks. Qwen2.5-VL-7B obtains 0.00 mAP on VidOR. STTran and TEMPURA reach 48.09 and 64.77 mR@20 on AG, but the prefix recognizer and delayed reference have not been evaluated on AG, so these rows cannot yet be compared with them. Until it is filled, the experiments do not establish that predictive modeling improves on the delayed reference under an equal observation budget.

Independent reference heads. Table 1(d) compares two heads trained independently on the same proposals. The real-future delayed head exceeds the prefix-only head by +0.14 pp mAP, with a paired-video 95% interval of [+0.06, +0.24], and raises CD-15 AP from 3.14 to 3.63.

4.3 Ablation Study on VidOR

Table 2 examines future-model capacity, retrieval supervision, candidate evidence, and separation gating, with Gate4 kept intact throughout. Every variant stays within 0.03 pp of the full model. Using a single Gaussian, three candidates, no retrieval loss, or permuted candidate evidence lowers mAP by 0.022 to 0.029 pp. Opening the separation gate or computing separation over the full mixture has almost no effect, which suggests that the gate rarely closes on VidOR. These differences are an order of magnitude smaller than the bootstrap interval of the main gain and come from single runs, so they do not isolate the contribution of individual components.

The permuted-evidence control is the most informative. It retains 91% of the full model's gain over the action expert, which indicates that most of the current improvement does not depend on candidate-specific futures. The VidVRD controls in Appendix A show the same pattern, with a random predictor and an unconditioned rollout within 0.02 pp of the full model.

Table 2: VidOR ablations (%). All RIVET rows use complete Gate4 and conditional fusion. Capacity and loss variants are retrained; evidence and separation controls replay the trained predictor. Δ is the mAP change from the full model. Single runs.

Variant	mAP	Δ	R@50	R@100
Action expert	11.985	−0.239	11.499	14.564
Gate4 only, no future branch	12.050	−0.174	11.453	14.521
RIVET + Gate4 ($M = 2$, $K = 10$)	12.224	–	11.436	14.505
$M = 1$ (single Gaussian)	12.202	−0.022	11.426	14.508
$K = 3$	12.195	−0.029	11.419	14.505
Without retrieval loss	12.198	−0.026	11.419	14.508
Permuted candidate evidence	12.202	−0.022	11.416	14.488
Separation gate always open	12.223	−0.001	11.429	14.505
Full-mixture separation	12.224	0.000	11.436	14.505

4.4 Qualitative Temporal Analysis

Figure 3 audits two VidOR development sequences. In the adult–cup example, the expert ranks the comparator release above the annotated hold by 0.0004, and the no-future control leaves this near tie unchanged. Matching the candidate futures against the observed continuation raises hold to 0.98 and yields a margin of +0.28. In the adult–bicycle example, the no-future control prefers drive over the annotated ride by 0.12, and future matching reverses the margin to +0.19. In both cases the comparator score stays fixed, so the correction comes from raising the annotated relation rather than suppressing the competitor. Failure cases appear in Appendix.

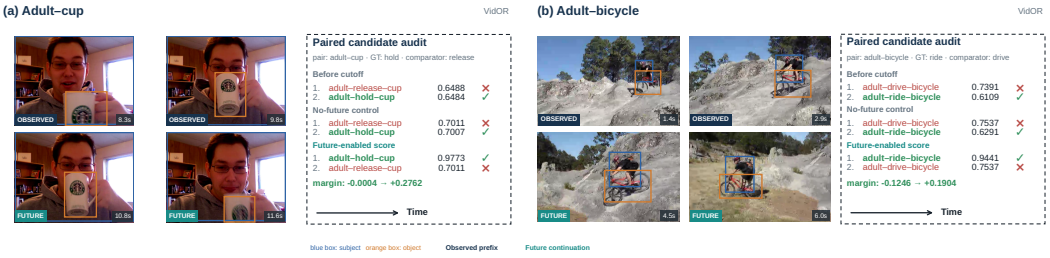

Figure 3: Future matching on two VidOR development sequences. (a) Adult–cup, ground truth hold versus release. (b) Adult–bicycle, ground truth ride versus drive. Frames show the observed prefix and continuation with subject (blue) and object (orange) boxes. Side panels give the candidate margin before the cutoff, under a no-future control, and with RIVET.

5 Conclusion

We studied streaming HOI detection under bounded lag, where a predicate can be revised once after a fixed delay while localization stays frozen. We proposed a matched-lag protocol that gives all methods the same continuation, and RIVET, which matches predicate-conditioned future predictions against the observed continuation and calibrates the revision to control expected degradation. Experiments show that the continuation carries useful predicate information and that RIVET gives small but consistent gains on VidOR. Our analyses also show that simulated futures still differ clearly from real ones, which limits the benefit of candidate-specific prediction. Closing this gap and building benchmarks with more consequence-defined predicates are promising directions for future work.

AI USE STATEMENT

During manuscript preparation, generative AI tools assisted in refining the wording and improving readability. The methodology details any use of these tools as part of the research itself. All AI-assisted revisions were reviewed by the authors, who retain full responsibility for the accuracy and integrity of the final manuscript.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

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 preprint arXiv:2506.09985*, 2025.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15791–15801, 2025.

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

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), 2021.

Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2Act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *Proceedings of the European Conference on Computer Vision*, 2024.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning*, 2024.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. TemporalBench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. VideoLLM-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Meng-Jiun Chiou, Chun-Yu Liao, Li-Wei Wang, Roger Zimmermann, and Jiashi Feng. ST-HOI: A spatial-temporal baseline for human-object interaction detection in videos. In *Proceedings of the Workshop on Intelligent Cross-Data Analysis and Retrieval (ICMR Workshops)*, 2021.

Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. Online action detection. In *Proceedings of the European Conference on Computer Vision*, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Pei Geng, Jian Yang, and Shanshan Zhang. HORP: Human-object relation priors guided HOI detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25325–25335, 2025.

Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Yupeng Hu, Changxing Ding, Chang Sun, Shaoli Huang, and Xiangmin Xu. Bilateral collaboration with large vision-language models for open vocabulary human-object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20126–20136, 2025.

Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10236–10247, 2020.

Xinjie Jiang, Chenxi Zheng, Xuemiao Xu, Bangzhen Liu, Weiying Zheng, Huaidong Zhang, and Shengfeng He. VrdONE: One-stage video visual relation detection. In *Proceedings of the ACM International Conference on Multimedia*, pp. 1437–1446, 2024.

Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

Hema S. Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38 (1):14–29, 2016.

Huy Le, Nhat Chung, Tung Kieu, Jingkang Yang, and Ngan Le. UNO: Unifying one-stage video scene graph generation via object-centric visual representation learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2797–2807, 2026.

Ting Lei, Shaofeng Yin, and Yang Liu. Exploring the potential of large foundation models for open-vocabulary HOI detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16657–16667, 2024.

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Alexander C. Li, Ananya Kumar, and Deepak Pathak. Generative classifiers avoid shortcut solutions. In *International Conference on Learning Representations*, 2025a.

Mengtian Li, Yu-Xiong Wang, and Deva Ramanan. Towards streaming perception. In *Proceedings of the European Conference on Computer Vision*, 2020.

Yanjun Li, Zhaoyang Li, Honghui Chen, and Lizhi Xu. Unbiased video scene graph generation via visual and semantic dual debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Sayak Nag, Kyle Min, Subarna Tripathi, and Amit K. Roy-Chowdhury. Unbiased scene graph generation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Zhanzhong Pang, Fadime Sener, and Angela Yao. Context-enhanced memory-refined transformer for online action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8700–8710, 2025.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Rohith Peddi, Saksham Singh, Saurabh, Parag Singla, and Vibhav Gogate. Towards scene graph anticipation. In *Proceedings of the European Conference on Computer Vision*, 2024.

Rohith Peddi, Saurabh, Ayush Abhay Shrivastava, Parag Singla, and Vibhav Gogate. Towards unbiased and robust spatio-temporal scene graph generation and anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8648–8657, 2025.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video LLMs with active real-time interaction via disentangled perception, decision, and reaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24045–24055, 2025.

Dominik Schmidt and Minqi Jiang. Learning to act without actions. In *International Conference on Learning Representations*, 2024.

Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.

Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *Proceedings of the ACM International Conference on Multimedia*, 2017.

Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the International Conference on Multimedia Retrieval*, pp. 279–287, 2019.

Tomáš Souček, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. Look for the change: Learning object states and state-modifying actions from untrimmed web videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Guan Wang, Zhimin Li, Qingchao Chen, and Yang Liu. OED: Towards one-stage end-to-end dynamic scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27938–27947, 2024.

Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. StreamBridge: Turning your offline video large language model into a proactive streaming assistant. In *Advances in Neural Information Processing Systems*, 2025a.

Liwei Wang, Yanduo Zhang, Tao Lu, Fang Liu, Huiqin Zhang, Jiayi Ma, and Huabing Zhou. End-to-end entity-predicate association reasoning for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b.

Meng Wei, Long Chen, Wei Ji, Xiaoyu Yue, and Roger Zimmermann. In defense of clip-based video relation detection. *IEEE Transactions on Image Processing*, 33:2759–2769, 2024.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. iVideoGPT: Interactive VideoGPTs are scalable world models. In *Advances in Neural Information Processing Systems*, 2024a.

Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. VadCLIP: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024b.

Zihui Xue, Kumar Ashutosh, and Kristen Grauman. Learning object state changes in videos: An open-world perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18493–18503, 2024.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. In *International Conference on Learning Representations*, 2025.

Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. AntGPT: Can large language models help long-term action anticipation from videos? In *International Conference on Learning Representations*, 2024.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *International Conference on Machine Learning*, 2025.