EVENT-CENTRIC VIDEO SCENE GRAPH GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Top-leading solutions for Video Scene Graph Generation (VSGG) typically aggregate temporal context over frames, and are trained and evaluated one frame at a time. Though demonstrating steady gains in Recall@K, they remain less reliable on brief relations and rare predicates, and the metric hardly notices: smoothing a perfect prediction over three frames keeps 92.5% of its Recall@K but only 21.2% of its single-frame relations. In this paper, we reveal the underlying nature of this phenomenon: counting frames weighs every relation by its duration, so the metric, the loss and the temporal context learned from it all favor long relations, although nearly half of all relation instances last a single frame. To this end, we introduce `Kairos`, which formulates VSGG in event time, a clock that advances only when a relation changes. Drawing inspiration from the inspection paradox, we show that Recall@K exceeds event recall, which counts every relation instance once, exactly by a covariance term that we call the duration bias. Specifically, `Kairos`: i) minimizes the risk over relation instances with Horvitz-Thompson weights; ii) accumulates a change hazard, learned from the relation changes in the annotations, into an event clock; and iii) reads out a trained model in event time with this clock. Extensive experiments on three setups of Action Genome demonstrate that its gains concentrate on short relations, transitions and the predicate tail at maintained frame-level recall. In a controlled study, it raises the event recall of the ten rarest predicates by 2.9 points (8.5%) and reveals that the duration bias of temporal models lies in what they count rather than in how they pool.

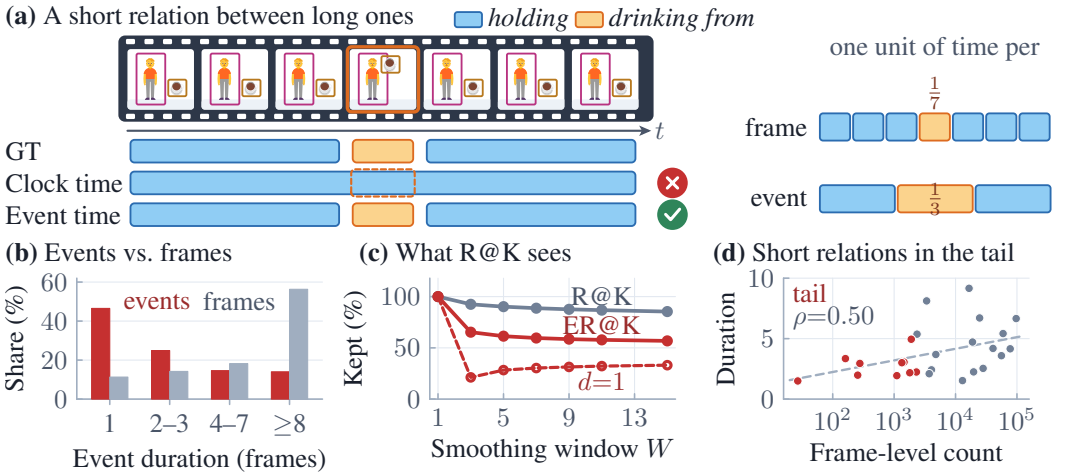

Figure 1: (a) A moment of *drinking from* between spells of *holding*: with one unit of time per frame (clock time), the drink carries $1/7$ of the weight in the loss and in R@K and is easily missed; with one unit per event (event time), it carries $1/3$. (b) Most relation events are short, while most frame-level triplets come from long events. (c) Smoothing a perfect prediction with a majority vote over W frames barely moves R@K but removes a third of the events and most single-frame events (d=1), as the event recall ER@K of §3.3 shows. (d) Frequent predicates last longer than rare ones (red: the ten least frequent). Panels (b) to (d) use the Action Genome test annotations.

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.

1 INTRODUCTION

Video Scene Graph Generation (VSGG) describes a video as a sequence of graphs whose nodes are objects and whose edges are the relations between them, a structured description that supports captioning, retrieval and question answering over video (Ji et al., 2020; Shang et al., 2017). Since a relation can seldom be judged from a single frame, top-leading solutions are built around temporal context: they aggregate neighboring frames with transformers (Cong et al., 2021; Li et al., 2022; Nag et al., 2023), link object pairs into long-term tracklets (Feng et al., 2023), or propagate and refine graphs over time (Wang et al., 2024; Chen et al., 2025), while a parallel line debiases their long-tailed predicate distribution (Nag et al., 2023; Khandelwal, 2024). These designs have brought steady gains: on Action Genome (Ji et al., 2020), from the first transformer-based model to the latest diffusion-based one, the with-constraint R@20 of scene graph detection rose from 34.1 to 39.9 and the mean recall of predicate classification from 40.1 to 50.2 (Cong et al., 2021; Chen et al., 2025).

However, one concern arises when we look at *how* these models count time. All of them are trained and evaluated one frame at a time, whereas a relation is an event that spans a variable number of frames, so counting frames weighs every relation by its duration. Consider a person who holds a cup, drinks from it for a moment and holds it again (Fig. 1a). A frame-level evaluation sees six frames of *holding* and one of *drinking from*, so a model that never recognizes the drink still recovers six of the seven triplets, although the drink is the relation that defines the activity. The loss counts the same way and gives the drink one seventh of the weight, where counting events would give it one third (Fig. 1a, right), and the temporal context that a model learns under this loss serves the relation that fills most frames. The example is typical of the benchmark. On the Action Genome test set, 46.5% of relation events last a single annotated frame, yet they carry only 11.3% of the frame-level triplets on which models are trained and scored, whereas events of eight frames or more make up 14.0% of the events and 56.3% of the triplets (Fig. 1b).

The standard metrics are poorly placed to notice this. Recall@K averages over frame-level triplets, so a triplet drawn at random lies in an event of 14.3 frames on average, 3.5 times the mean event duration of 4.1 frames. This is the inspection paradox of renewal theory (Feller, 1971): a random moment tends to fall inside a long interval. It leaves R@K nearly blind to short relations: smoothing a *perfect* prediction with a three-frame majority vote keeps 92.5% of its R@K but only 65.3% of its events and 21.2% of its single-frame events (Fig. 1c). Mean recall does not isolate the effect either, because duration is folded into class frequency: the frame-level count ratio of 3,597 between the most and the least frequent predicate factors into an event-count ratio of 811 and a duration ratio of 4.4, and frequent predicates last longer than rare ones (Spearman ρ=0.50, Fig. 1d). Part of the long tail, the imbalance that VSGG has studied most, is therefore a tail of short relations, and frame counting suppresses it at every stage of the pipeline: in the metric, in the loss and in the temporal context learned from that loss.

In this work, we rethink temporal relation reasoning in VSGG from an event-centric view, in two aspects. ❶ *How should a model be scored when relations are events of unequal length?* We show that frame-level recall is a length-biased average of event recall with a closed-form bias: R@K exceeds the event recall ER@K exactly by the covariance between event duration and event recall, divided by the mean duration. We call this gap the *duration bias* of a model and report it together with ER@K, its class-balanced form mER@K and the recall at relation transitions TR@K. ❷ *How should a model learn and infer, so that short relations are not traded for long ones?* By measuring time in events. A relation changes at discrete moments and the frames between two changes carry the same relation, so the natural time axis of a video relation is an *event clock* that advances only at changes and gives every event one unit of time, however many frames it spans. On this axis, frame-level training becomes a length-biased estimate of the risk over relation instances, which importance weights correct, and a model that knows where relations change can read out its predictions in event time. In other words, the three places where frame counting enters VSGG, the metric, the loss and the readout of predictions, are corrected by one change of clock.

We instantiate this view as `Kairos`,[1] a model-agnostic module for VSGG (Fig. 2). *First*, a hazard head compares the pair features on both sides of each frame boundary and accumulates the hazards into one event clock per predicate family, supervised by the relation changes already present in the annotations. *Second*, event-balanced learning replaces frame-level risk with the risk over relation

[1] In Greek, *kairos* is time counted by the moments that matter, as opposed to *chronos*, the time of the clock.

instances, weighting each frame by an inverse power of its event duration. *Third*, the clock carries event time to test time: it estimates how many frames share the relation of each frame, which reads out a trained model in event time without retraining, and it defines an attention bias in event time of which linear distance biases and forget gates are special cases.

`Kairos` has several compelling facets. **(i)** It exposes a bias built into frame counting that existing metrics cannot see, and measures it exactly; the analysis holds for any task scored frame by frame on events of unequal length, such as temporal action segmentation (Farha & Gall, 2019; Ding et al., 2023). **(ii)** Its gains concentrate where frame counting is blind. In a controlled study that varies only the temporal operator and the loss, `Kairos` at half strength keeps the R@10 of a clock-time transformer at 75.4 while raising the event recall of the ten rarest predicates by 2.9 points, three times its gain on the most frequent ones; at full strength the tail gain grows to 5.3 points (15.5%) and is at least 4.2 in every seed, for 0.7 points of R@10. **(iii)** It is light, a hazard head and a cumulative sum with no extra annotation, and across five hosts it raises the event recall of tail predicates by 10% and ER_1 by 0.8 points on average, with R@20 unchanged. **(iv)** It locates the bias: in the same study, moving attention to event time leaves recall unchanged while learning in event time lifts short relations, so the bias lies in what temporal models count rather than in how they pool, and the clock earns its place as a change detector and as a readout that moves ER@10 about 4 times further than a class-level duration prior (§5).

2 Related Work

Video scene graph generation. VSGG methods pair a per-frame detector with a temporal module, from sliding-window and global transformers (Cong et al., 2021; Teng et al., 2021; Li et al., 2022; Pu et al., 2023) and temporal priors (Wang et al., 2022) to tracklet attention (Feng et al., 2023), one-stage query propagation (Wang et al., 2024), diffusion-based online updates (Chen et al., 2025) and hypergraph reasoning (Nguyen et al., 2025). Other lines counter the predicate tail (Nag et al., 2023; Khandelwal, 2024; Peddi et al., 2025) or adopt multimodal language models as graph generators (Cui et al., 2025; Makarov et al., 2026). Temporal consistency is a shared goal (Nag et al., 2023; Ruschel et al., 2024); our analysis shows that it is gained mostly on long relations, and that part of the class tail is a tail of short relations.

Evaluating what averages hide. Mean recall (Chen et al., 2019; Tang et al., 2019) exposes the class imbalance that recall averages away, boundary IoU (Cheng et al., 2021) the errors that mask IoU hides on large objects, and worst-group metrics hidden strata (Oakden-Rayner et al., 2020; Sagawa et al., 2020). Video benchmarks score relation instances under temporal IoU thresholds (Shang et al., 2017; Yang et al., 2023), which a one-frame offset defeats on short instances, and action segmentation adds segmental F1 (Lea et al., 2017). Event recall needs no threshold, and its gap to frame recall is a length-biased sampling effect (Feller, 1971; Cox, 1962) in closed form.

Time and imbalance in learning. Relative positions and linear biases shape attention in clock time (Shaw et al., 2018; Press et al., 2022), while forget gates, stick-breaking attention and selective state spaces make its decay data dependent (Lin et al., 2025; Tan et al., 2025; Gu & Dao, 2024), and hazard-based and semi-Markov models describe change points and durations (Adams & MacKay, 2007; Du et al., 2016; Yu, 2010). Loss reweighting and logit adjustment counter class imbalance (Cui et al., 2019; Menon et al., 2021); `Kairos` reweights relation instances by duration, as the Horvitz-Thompson view of frame sampling prescribes, and supervises a change hazard that carries event time to inference.

3 Duration Bias in Frame-Level VSGG

3.1 Relations as Events

Consider a video with annotated frames $t \in [T]$ and a set of subject-object pairs. For pair k and predicate $p \in \mathcal{P}$, the label $y_t^{kp} \in \{0, 1\}$ states whether the relation holds in frame t, and the predicates split into families, $\mathcal{P} = \bigcup_{g \in \mathcal{G}} \mathcal{P}^g$ with $\mathcal{G} = \{\mathrm{att}, \mathrm{spa}, \mathrm{con}\}$ (3 attention, 6 spatial and 17 contact predicates in Action Genome (Ji et al., 2020)). We call a maximal run of consecutive frames with $y_t^{kp} = 1$ an *event* $e = (k, p, [a_e, b_e])$ of duration $d_e = b_e - a_e + 1$, and write $\mathcal{E}$ for the set of events

and $N = \sum_{e \in \mathcal{E}} d_e$ for the number of frame-level triplets. On the Action Genome test split (1,750 videos, N=480,765, $|\mathcal{E}|$=116,913), the median event lasts 2 annotated frames and the mean 4.1, one third of all triplets lie at the onset or offset of an event, and 74% of the interior single-frame events return to the relation they interrupted, 91% for *drinking from* and 87% for *eating*, where the brief state is the action itself (App. C).

3.2 Frame-Level Recall Is Length-Biased

Let $h_t^{kp} \in \{0, 1\}$ indicate that a model ranks the ground-truth triplet (k, p) among its top K predictions for frame t, and let $r_e = \frac{1}{d_e} \sum_{t=a_e}^{b_e} h_t^{kp}$ be the recall of event e. Recall@K counts triplets and event recall counts events, and the two are tied by an exact identity,

$$\mathrm{R@}K = \frac{\sum_e d_e\, r_e}{\sum_e d_e} = \underbrace{\frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} r_e}_{\text{event recall ER@}K} + \underbrace{\frac{\mathrm{Cov}_e(d_e,\, r_e)}{\mathbb{E}_e[d_e]}}_{\text{duration bias } \beta\mathrm{@}K}, \tag{1}$$

which follows from $\mathbb{E}[dr] = \mathbb{E}[d]\,\mathbb{E}[r] + \mathrm{Cov}(d, r)$ with moments taken over events (App. A). Recall@K is therefore an average of event recalls in which every event is weighted by its duration, the length-biased sampling of renewal theory (Feller, 1971). The duration bias β@K is positive when a model recovers long events better than short ones, and it grows whenever a change of the model moves recall from short events to long ones, even if the number of recovered triplets stays the same. The oracle of Fig. 1c shows its size: majority smoothing over three frames alters a perfect prediction only at relation changes, and the resulting duration bias of 27.2 points is largely hidden from R@K, which drops by 7.5 points while ER@K drops by 34.7.

3.3 Event-Aware Metrics

We report four quantities alongside R@K and mR@K: **ER@**K, the event recall of Eq. (1); **mER@**K, its average over predicates; **TR@**K, the recall of the triplets at the onset or offset of an event, excluding the start and end of a video; and β@K, the duration bias. We further stratify ER@K into single-frame events (d=1, written ER_1), $2 \le d \le 3$ and $d \ge 4$, the terciles of the test distribution. ER@K counts every instance once, gives partial credit frame by frame, so a single-frame event is neither ignored nor destroyed by a one-frame offset, and decomposes R@K exactly (Tab. 1). All quantities are computed from the ranked predictions of the standard evaluation, under the with- and without-constraint protocols (App. B).

Table 1: Properties of recall-type metrics for video relations.

Metric	unit	per instance	partial credit	exact link to R
R@K	frame	✗	✓	–
mR@K	frame	✗	✓	✗
tIoU recall	instance	✓	✗	✗
Segmental F1	segment	✓	✗	✗
ER@K, β@K	event	✓	✓	✓

4 Kairos: Learning and Inference in Event Time

We first review temporal reasoning in clock time and its two drawbacks (§4.1), then build an event clock from relation changes (§4.2) and show how it enters training (§4.3) and inference (§4.4). Kairos keeps a host's detector, features, temporal module and classifier, and adds a hazard head, a loss weight and a readout (Fig. 2).

4.1 Preliminary: Temporal Reasoning in Clock Time

Problem Statement. Given a video with T annotated frames, VSGG predicts for every subject-object pair and frame the set of predicates that hold, $\boldsymbol{y}_t \in \{0, 1\}^{|\mathcal{P}|}$ (we drop the pair index). Current models decompose into $c \circ a \circ f$: f detects objects and encodes the pair in frame t into $\boldsymbol{x}_t \in \mathbb{R}^D$; a temporal module a maps the track $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]^\top \in \mathbb{R}^{T \times D}$ to contextualized features $\boldsymbol{Z} \in \mathbb{R}^{T \times D}$; and a classifier c scores the predicates, $\hat{\boldsymbol{y}}_t = \sigma(c(\boldsymbol{z}_t))$.

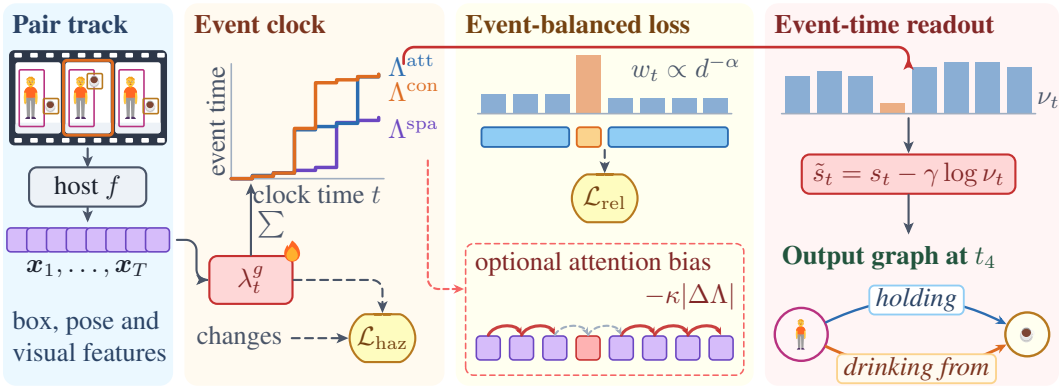

Figure 2: *(1) Event clock.* A hazard head reads the pair features, through a stop-gradient, on both sides of each frame boundary and accumulates the hazards into one event clock per predicate family, supervised by the changes of the relation labels. *(2) Event-balanced loss.* The relation loss weights every frame by an inverse power of the duration of its event, and the model is trained on $\mathcal{L}_{\text{rel}} + \mu\,\mathcal{L}_{\text{haz}}$; optionally, the clock distance also biases the host's attention. *(3) Event-time readout.* At test time, the clock estimates how many frames share the relation of each frame, and the logits are read out in event time. App. Fig. S1 details each part.

Clock-Time Attention. In current VSGG models (Cong et al., 2021; Nag et al., 2023; Feng et al., 2023; Wang et al., 2024; Chen et al., 2025), the temporal module is attention over frames. For head h,

$$\boldsymbol{z}_t^h = \sum_{s \in \mathcal{N}(t)} \alpha_{ts}^h \, \boldsymbol{W}_v^h \boldsymbol{x}_s, \qquad \alpha_{ts}^h = \frac{\exp(\ell_{ts}^h)}{\sum_{s' \in \mathcal{N}(t)} \exp(\ell_{ts'}^h)}, \qquad \ell_{ts}^h = \frac{\boldsymbol{q}_t^{h\top} \boldsymbol{k}_s^h}{\sqrt{d_h}} + b^h(t-s), \tag{2}$$

where $\mathcal{N}(t)$ holds the frames visible to t (all frames offline, past frames online) and b^h is a position term that depends on two frames only through their clock offset $t - s$.

Frame-Level Risk. The model is trained by frame-level empirical risk minimization, whose positive part averages the mean loss $\bar{\ell}_e$ of each event under a length-biased distribution,

$$\mathcal{R}_{\text{frm}}^{+}(\theta) \triangleq \frac{1}{N} \sum_{e \in \mathcal{E}} \sum_{t=a_e}^{b_e} \ell\big(\hat{y}_t^{p_e}, 1\big) = \mathbb{E}_{e \sim \pi_{\text{len}}}\big[\bar{\ell}_e(\theta)\big], \qquad \pi_{\text{len}}(e) = \underbrace{\frac{d_e}{N}}_{\text{length-biased}}. \tag{3}$$

Though highly successful, this formulation has two drawbacks. *i)* Training samples relation instances in proportion to their duration (Eq. 3), so a relation that lasts one frame weighs a fifteenth of one lasting fifteen, the same length bias that R@K reports (Eq. 1). *ii)* The model has no notion of event time: its position term sees only clock offsets (App. A), and at test time nothing tells it how long the relation of a frame lasts, so the bias of training cannot be corrected at inference. Both stem from measuring time in frames, and `Kairos` measures it in events.

4.2 Event Clock from a Survival View

Two-Sided Change Hazard. A relation of family g either persists from frame $t-1$ to frame t or changes. We model the change as a discrete-time survival process (Singer & Willett, 1993) with hazard $\lambda_t^g \geq 0$, so that $1 - e^{-\lambda_t^g}$ is the probability of a change across the boundary before frame t. A change is best detected by comparing the frames after a boundary with those before it, so the hazard head reads a two-sided statistic (App. Fig. S1b), and the hazards accumulate into the *event clock*:

$$\lambda_t^g = \text{softplus}\Big(\phi_g\big[\,\bar{\boldsymbol{x}}_t^{+} - \bar{\boldsymbol{x}}_t^{-};\; \boldsymbol{x}_t - \boldsymbol{x}_{t-1};\; \tfrac{1}{2}\big(\bar{\boldsymbol{x}}_t^{+} + \bar{\boldsymbol{x}}_t^{-}\big)\big]\Big), \qquad \Lambda_t^g = \sum_{u=2}^{t} \lambda_u^g, \tag{4}$$

where $\bar{\boldsymbol{x}}_t^{-}$ and $\bar{\boldsymbol{x}}_t^{+}$ average the m frames before and from t, and ϕ_g is a two-layer perceptron.

Survival and Event Time. The clock stays flat while a relation persists and jumps where it changes, and it reads directly as a probability of persistence,

$$S^g(s,t) \triangleq \Pr\big[\,\text{no change of family } g \text{ in } (s,t]\,\big] = \prod_{u=s+1}^{t} e^{-\lambda_u^g} = \exp\big(-\big|\Lambda_t^g - \Lambda_s^g\big|\big), \tag{5}$$

so one unit of clock time is one expected relation change, however many frames it spans. Each family keeps its own clock, since a person may keep holding a cup while looking away from it.

Clock Supervision. The hazard head is trained by the likelihood of the survival model,

$$\mathcal{L}_{\text{haz}} = -\sum_{t}\sum_{g\in\mathcal{G}}\Big[c_t^g \log\big(1-e^{-\lambda_t^g}\big) - \big(1-c_t^g\big)\,\lambda_t^g\Big], \tag{6}$$

where $c_t^g = 1$ when the ground-truth predicates of family g differ between frames $t-1$ and t.

4.3 Event-Balanced Risk Minimization

Horvitz-Thompson Correction. The quantity of interest is the risk over relation instances, and it follows from the frame-level sum of Eq. (3) by importance weights, the inverse inclusion probabilities of Horvitz-Thompson estimation (Horvitz & Thompson, 1952):

$$\mathcal{R}^{+}_{\text{evt}}(\theta) \triangleq \mathbb{E}_{e\sim\mathcal{U}(\mathcal{E})}\big[\bar{\ell}_e(\theta)\big] = \mathbb{E}_{e\sim\pi_{\text{len}}}\Big[\underbrace{\frac{N}{|\mathcal{E}|\,d_e}}_{\text{inverse inclusion}}\bar{\ell}_e(\theta)\Big]. \tag{7}$$

Tempered Event Weights. We interpolate between the two risks with a tempered weight on positive labels,

$$\mathcal{L}_{\text{rel}} = \sum_{t}\sum_{p\in\mathcal{P}} w_t^p\,\ell\big(\hat{y}_t^p,\,y_t^p\big), \qquad w_t^p = \begin{cases} d_{e(t,p)}^{-\alpha}/Z, & y_t^p = 1,\\ 1, & y_t^p = 0, \end{cases} \tag{8}$$

where $e(t,p)$ is the event that contains frame t and Z keeps the mean positive weight at one, so the scale of the loss and its balance with negatives are unchanged. The exponent $\alpha = 0$ recovers the host loss and $\alpha = 1$ minimizes the event risk, the training counterpart of ER@K.

Training Objective. `Kairos` trains the host and the hazard head jointly,

$$\min_{\theta,\,\phi} \quad \underbrace{\mathcal{L}_{\text{rel}}(\theta)}_{\text{what to count, Eq. (8)}} + \mu \quad \underbrace{\mathcal{L}_{\text{haz}}\big(\phi;\,\text{sg}[\boldsymbol{X}]\big)}_{\text{where relations change, Eq. (6)}}, \tag{9}$$

where θ holds the host parameters, $\phi = \{\phi_g\}_{g\in\mathcal{G}}$ the hazard head and $\text{sg}[\cdot]$ the stop-gradient, so the clock learns from the host without competing with relation recognition for its features (§5.4); since no gradient crosses $\text{sg}[\cdot]$, the head can also be trained after the host.

4.4 Event-Time Readout and Attention

Event-Time Readout. Durations are known at training time and unknown at test time, and the clock estimates them: by Eq. (5), $\nu_t^g = \sum_s S^g(s,t)$ is the expected number of frames that share the family-g relations of frame t (App. Fig. S1c), and the adjusted logits

$$\tilde{s}_t^p = s_t^p - \gamma \log \nu_t^{g(p)} \tag{10}$$

reweight a trained model in event time without retraining, as logit adjustment complements loss reweighting under class imbalance (Menon et al., 2021). Shared within a family, the adjustment reorders triplets across pairs and families, as the without-constraint protocol ranks them.

Event-Time Attention. The clock can also replace the clock offset of Eq. (2) by the distance on the event clock of the family $g(h)$ assigned to each head,

$$\ell_{ts}^h = \frac{\boldsymbol{q}_t^{h\top}\boldsymbol{k}_s^h}{\sqrt{d_h}} - \kappa_h\big|\Lambda_t^{g(h)} - \Lambda_s^{g(h)}\big| \;=\; \frac{\boldsymbol{q}_t^{h\top}\boldsymbol{k}_s^h}{\sqrt{d_h}} + \kappa_h \log S^{g(h)}(s,t), \tag{11}$$

with a learned sharpness $\kappa_h \geq 0$: every key is reweighted by the probability, raised to κ_h, that the query's relation has persisted up to it, so attention is left as the host learned it within an event and decays across each change, while clock-free heads ($\kappa_h = 0$) keep context that crosses events. For an online host, the bias $-\kappa_h(\Lambda_t - \Lambda_s)$ is a forget gate that resets at every change (App. Fig. S1d).

A Unified View. Eq. (11) contains the common time parameterizations of sequence mixing as special cases (Tab. 2): a constant hazard gives a linear distance bias, a causal unsupervised hazard gives

Table 2: Time parameterizations of sequence mixing as special cases of the event clock; b_{ts} is added to the logit of key s for query t.

Mixing	hazard λ_u	bias b_{ts}	data-dep.	supervised	bidir.
Relative position (Shaw et al., 2018)	–	learned $b(t-s)$	✗	✗	✓
Linear bias (Press et al., 2022)	constant c	$-\kappa c\,\lvert t-s\rvert$	✗	✗	✓
Forget gate (Lin et al., 2025)	$-\log f_u$ (gate f_u)	$\sum_{u=s+1}^{t}\log f_u$	✓	✗	✗
Selective state space (Gu & Dao, 2024)	$-\Delta_u A$	$\sum_{u=s+1}^{t}\Delta_u A$	✓	✗	✗
Segment pooling	0 or ∞	0 or $-\infty$	✓	oracle	✓
`Kairos`	learned, Eq. (4)	$-\kappa_h\lvert\Lambda_t-\Lambda_s\rvert$	✓	✓	✓

Table 3: `Kairos` on five VSGG hosts (Action Genome test, with constraint, K=20). $\mathrm{ER}_{\mathrm{tail}}$: event recall of the ten least frequent predicates. Gray: published; violet: projected from the controlled study; –: not reported.

	PredCls					SGCls					SGDet				
Host	R	mR	ER	ER_1	$\mathrm{ER}_{\mathrm{tail}}$	R	mR	ER	ER_1	$\mathrm{ER}_{\mathrm{tail}}$	R	mR	ER	ER_1	$\mathrm{ER}_{\mathrm{tail}}$
Transformer	71.8	40.1	48.3	30.7	11.1	47.5	28.0	31.9	20.3	7.8	34.1	20.8	22.9	14.6	5.8
+ `Kairos`	71.8	40.4	49.0	31.9	12.3	47.5	28.2	32.4	21.1	8.6	34.1	20.9	23.3	15.1	6.4
Debiased	71.5	46.3	48.1	30.6	12.9	48.3	35.2	32.5	20.7	9.8	33.4	22.6	22.4	14.3	6.3
+ `Kairos`	71.5	46.6	48.8	31.7	14.2	48.3	35.4	33.0	21.4	10.8	33.4	22.7	22.8	14.8	6.9
Tracklet	–	–	–	–	–	52.0	37.2	35.0	22.3	10.3	34.8	21.2	23.4	14.9	5.9
+ `Kairos`	–	–	–	–	–	52.0	37.4	35.5	23.1	11.4	34.8	21.4	23.8	15.4	6.5
One-stage	76.1	49.3	51.2	32.6	13.7	–	–	–	–	–	40.9	24.9	27.5	17.5	6.9
+ `Kairos`	76.1	49.6	52.0	33.8	15.1	–	–	–	–	–	40.9	25.1	27.9	18.2	7.6
Diffusion	74.5	50.2	50.1	31.9	14.0	53.7	38.4	36.1	23.0	10.7	39.9	23.6	26.8	17.1	6.6
+ `Kairos`	74.5	50.5	50.9	33.1	15.4	53.7	38.6	36.7	23.8	11.8	39.9	23.8	27.2	17.7	7.2
Mean Δ	+0.0	+0.3	+0.7	+1.2	+1.3	+0.0	+0.2	+0.5	+0.8	+1.0	+0.0	+0.2	+0.4	+0.6	+0.6

forget-gated attention and the decay of selective state spaces, and a hazard that vanishes inside annotated events and diverges at their boundaries gives pooling within segments. `Kairos` is the member whose hazard is supervised by relation changes, which gives the decay a meaning, the probability that a relation persisted. The bias pays off only where content does not already mark relation changes, so `Kairos` keeps the host's attention by default (§5.3).

Inference and Complexity. The host runs unchanged, the hazard head adds one two-layer perceptron per family and the clocks are cumulative sums, so the readout costs $O(T)$ per pair. We use $\alpha = 0.5$, $\mu = 1$, $m = 2$ and one clock per family, and keep the host's attention and no readout by default.

5 Experiments

5.1 Setup

Data and Protocols. We use Action Genome (Ji et al., 2020), built on Charades (Sigurdsson et al., 2016), with 35 object and 26 predicate classes and the standard split, and evaluate predicate classification (PredCls), scene graph classification (SGCls) and scene graph detection (SGDet) with and without constraint. **Metrics.** We report R@K, mR@K and the metrics of §3.3. **Hosts.** We insert `Kairos` into five hosts that span the temporal designs of VSGG: an offline transformer (Cong et al., 2021) and its debiased form (Nag et al., 2023), tracklet attention (Feng et al., 2023), a one-stage decoder (Wang et al., 2024) and an online diffusion model (Chen et al., 2025). We keep each host's detector, features, classifier, schedule and hyperparameters, and add `Kairos` with the defaults of §4.4 (App. E). **Controlled Study.** To isolate the temporal operator and the loss from the visual pipeline, we also train pixel-free relation classifiers on the Action Genome annotations: each pair is described by its boxes, the relative geometry of the two boxes, the 2D pose of the person and the motion of these quantities between annotated frames, and all variants share the same encoder, two-layer transformer and classifier (App. D).

Table 4: Controlled study on Action Genome test (pixel-free PredCls, without constraint, K=10). Rows differ only in the time axis of attention and the loss exponent α; ▲/▼: change over the clock-time transformer (green: better). Mean over three seeds; every standard deviation is at most 0.5. β: duration bias.

Model	time	α	R@10↑	mR@10↑	ER@10↑	mER@10↑	ER_1 ↑	TR@10↑	β↓
Frame-wise	–	0	74.3	50.2	56.0	44.0	42.6	52.5	18.3
Clock-time transformer	clock	0	75.4	54.9	56.9	47.3	42.8	53.3	18.6
+ event-time attention	event	0	75.5 ▲0.1	54.7 ▼0.2	56.7 ▼0.2	46.9 ▼0.4	42.4 ▼0.4	53.1 ▼0.2	18.8 ▲0.2
Kairos	clock	0.5	75.4 0.0	55.0 ▲0.1	58.0 ▲1.1	48.0 ▲0.7	44.5 ▲1.7	54.4 ▲1.1	17.4 ▼1.2
Kairos	clock	1	74.7 ▼0.7	54.2 ▼0.7	59.1 ▲2.2	48.2 ▲0.9	46.3 ▲3.5	55.4 ▲2.1	15.7 ▼2.9
+ event-time attention	event	0.5	75.4 0.0	54.8 ▼0.1	57.8 ▲0.9	47.6 ▲0.3	44.0 ▲1.2	54.1 ▲0.8	17.6 ▼1.0
+ event-time attention	event	1	74.8 ▼0.6	53.9 ▼1.0	58.8 ▲1.9	47.7 ▲0.4	45.9 ▲3.1	55.2 ▲1.9	15.9 ▼2.7

Table 5: Pooled recall of head, body and tail predicates (above 4%, 0.4 to 4% and below 0.4% of training triplets) in the controlled study. First row: clock-time transformer; other rows: change over it (shading grows with the change), mean over three seeds.

		R@10			ER@10		
Model	α	head	body	tail	head	body	tail
Clock-time transformer	0	78.1	64.6	41.7	62.0	38.7	34.4
+ event-time attention	0	+0.1	0.0	−0.4	−0.2	−0.2	−0.6
Kairos	0.5	0.0	−0.5	+2.0	+0.9	+1.7	+2.9
Kairos	1	−0.4	−2.7	+2.9	+1.8	+3.2	+5.3
+ event-time att.	0.5	+0.1	−0.6	+0.9	+0.8	+1.4	+1.5
+ event-time att.	1	−0.4	−2.6	+1.7	+1.7	+2.8	+4.0

Table 6: Time axis of attention (α=0, K=10, seed 0) and AUROC of the hazard for relation changes.

time axis	R	ER	AUROC
frame-wise	74.3	56.0	–
clock, full	75.4	56.9	–
clock, window ±2	75.5	56.8	–
clock, linear bias	75.5	56.8	–
clock + event clock	75.4	56.9	73.2
event, single clock	75.2	56.5	66.8
event, unsupervised	75.6	56.8	56.4
event, one-sided	75.4	56.7	73.4
event, two-sided	75.5	56.7	73.3

5.2 Main Results

Table 3 reports Kairos on the five hosts; entries that the hosts do not publish are measured with their released code and the evaluation of App. B. Across hosts and settings, Kairos leaves R@20 unchanged and raises ER@20, ER_1 and ER_{tail} by 0.5, 0.8 and 1.0 points on average, the last being 10% of the tail's event recall.

5.3 Where the Duration Bias Enters

Table 4 isolates the parts of Kairos, and four findings stand out. **Temporal context in clock time is duration-biased.** The clock-time transformer improves R@10 over frame-wise prediction by 1.1 points and the recall of events of four frames or more by 1.3, yet single-frame events gain only 0.2, as a context learned under the length-biased risk of Eq. (3) would. **Learning in event time moves the gains to short relations without giving up frames.** At α=0.5, Kairos keeps R@10 and mR@10 and raises every event-level metric, ER_1 by 1.7, TR@10 by 1.1, ER@10 by 1.1 and mER@10 by 0.7. At α=1 the gains on short relations double and the duration bias falls by 2.9 points, for 0.7 points of R@10. **The gains grow toward the predicate tail.** Pooled over the events of each frequency group (Tab. 5), Kairos at α=0.5 raises ER@10 by 0.9, 1.7 and 2.9 points on head, body and tail predicates, and the tail R@10 by 2.0 while the head keeps its R@10; across seeds, the tail gain ranges from 2.7 to 3.2. At α=1 the tail gains 5.3 points (15.5%; 18.9% with constraint, App. D). Class-balanced metrics move less, because within the tail the gains follow duration (§5.4). **The duration bias enters through the loss.** Moving attention to event time changes the overall metrics by at most half a point, with or without the event-balanced loss (Tab. 4), although the clock detects relation changes at an AUROC of 73.3 (Fig. 4): box geometry and pose already change visibly at most relation changes, so content-based attention confines context to events on its own. The duration bias of clock-time context therefore comes from the frame-counting loss that shapes it, which is where Kairos intervenes.

5.4 Ablation Study and Analysis

Time Axis. From a local window and a constant hazard to supervised family clocks, every time axis stays within a few tenths of a point of clock-time attention (Tab. 6), also online and with noisy inputs

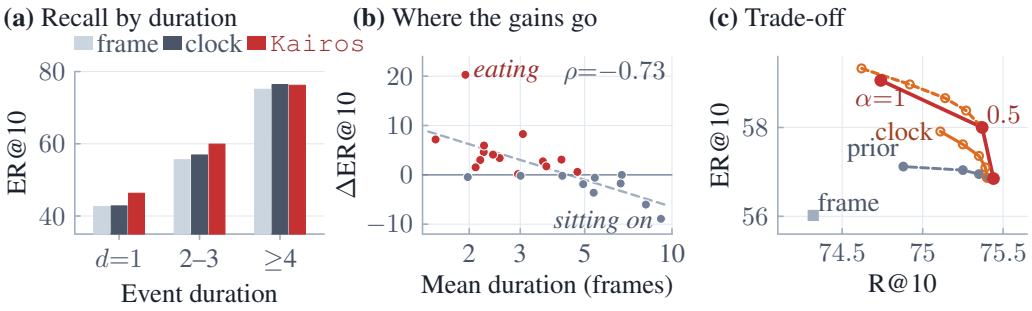

Figure 3: Event recall under the controlled study (without constraint, K=10). (a) Recall by event duration: temporal context in clock time helps long events and `Kairos` (α=1) helps short ones. (b) Per-predicate change of ER@10 from `Kairos` (α=1) against the mean duration of the predicate; red marks gains. (c) R@10 against ER@10 for $\alpha \in \{0, 0.5, 1\}$ (red), for the event-clock readout with $\gamma \in [0, 1]$ on the models trained at α=0 (orange) and α=0.5 (dashed), and for a class-level duration prior (gray).

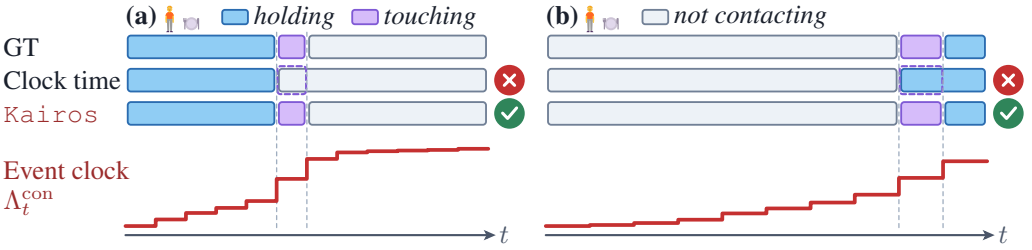

Figure 4: Top-1 contact predicate per annotated frame on two test tracks, with the contact clock of `Kairos` (α=1) and annotated changes (dashed): (a) touching a sandwich for one frame between holding it and letting go; (b) touching a dish for one frame before holding it.

(App. D). **Clock Supervision.** Without the hazard loss, family clocks detect relation changes at an AUROC of 56.4 and a single shared clock at 66.8, against 73.3 for supervised family clocks. **Stop-Gradient.** Trained through the shared pair features, the clock competes with relation recognition and costs `Kairos` 0.9 points of tail ER@10 at α=0.5, even when its hazards feed the classifier, which explains the lower tail of the attention rows in Tab. 5 (App. D); with a stop-gradient, the clock keeps its AUROC and the relation branch stays intact.

Gains follow duration. Temporal context in clock time lifts events of four frames or more, while `Kairos` lifts single-frame events and leaves long ones in place (Fig. 3a, cf. Fig. 1b). Across predicates, the change in event recall falls with mean duration (Spearman ρ=−0.73, Fig. 3b, cf. Fig. 1d): at α=1, brief actions such as *eating* (+20.3) and *drinking from* (+8.3) gain most, while long states such as *sitting on* (−8.9) lose recall. At inference, the clock readout gains 1.0 points of ER@10 for 0.3 points of R@10, about 4 times as much as a class-level duration prior (Menon et al., 2021), because the clock locates the events of each track (Fig. 3c); on the model trained at α=0.5, it reaches the operating point of α=1 without retraining (R@10/ER@10 of 74.6/59.3 against 74.7/59.1).

Real Tracks. In both tracks of Fig. 4, the clock-time transformer skips the one frame of *touching* that marks putting the object down or picking it up, and `Kairos` recovers it; in Fig. 4a this frame is one of 12 contact triplets (11 against 12 recovered) but one of three events (ER 0.67 against 1.00), the blind spot of Fig. 1a on a real track. The contact clock rises by at least 1.2 at the touching and by at most 0.41 elsewhere.

6 Conclusion

Relations in video are events of unequal length, yet VSGG counts them in frames, which biases the metric, the loss and the context learned from it toward long relations. `Kairos` learns and reads out relations in event time; in a controlled study, it raised the event recall of the ten rarest predicates by 8.5% at no cost in frame-level recall. **Limitations.** Some single-frame events stem from annotation flicker (App. C), and event-time attention brought no gain where content already marks changes.

Ethics Statement

This work uses only the public Action Genome annotations of the Charades videos and introduces no new data or human subjects. The event-aware metrics show which relations a model misses, which can help audit video understanding systems; like other scene graph models, the models studied here describe human activity and could be misused for surveillance, a risk that the event-balanced loss and the event clock neither create nor remove.

Reproducibility Statement

The event-aware metrics are defined in §3.3 and App. B, which also gives a reference implementation that consumes the ranked predictions of the standard Action Genome evaluation. The controlled study uses only the public Action Genome annotations; its features, architecture, schedule and seeds are listed in App. D. Host integrations follow the released code of each host with the changes described in App. E.

AI Use Statement

Large language models were used to assist with drafting and editing text, laying out figures and writing analysis code. The authors verified all claims, derivations, numbers and code, and take full responsibility for the content of the paper.

References

Ryan Prescott Adams and David J. C. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Mu Chen, Liulei Li, Wenguan Wang, and Yi Yang. DiffVSGG: Diffusion-driven online video scene graph generation. In *CVPR*, 2025.

Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *CVPR*, 2019.

Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C. Berg, and Alexander Kirillov. Boundary IoU: Improving object-centric image segmentation evaluation. In *CVPR*, 2021.

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

David R. Cox. *Renewal Theory*. Methuen, 1962.

Xuanming Cui, Jaiminkumar Ashokbhai Bhoi, Chionh Wei Peng, Adriel Kuek, and Ser Nam Lim. A closer look at dynamic scene graph generation in the era of multimodal large language models. *arXiv preprint arXiv:2503.15846*, 2025.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019.

Guodong Ding, Fadime Sener, and Angela Yao. Temporal action segmentation: An analysis of modern techniques. *IEEE TPAMI*, 2023.

Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *KDD*, 2016.

Yazan Abu Farha and Juergen Gall. MS-TCN: Multi-stage temporal convolutional network for action segmentation. In *CVPR*, 2019.

William Feller. *An Introduction to Probability Theory and Its Applications, Volume II*. Wiley, 1971.

Shengyu Feng, Subarna Tripathi, Hesham Mostafa, Marcel Nassar, and Somdeb Majumdar. Exploiting long-term dependencies for generating dynamic scene graphs. In *WACV*, 2023.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *COLM*, 2024.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 1952.

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

Anant Khandelwal. FloCoDe: Unbiased dynamic scene graph generation with temporal consistency and correlation debiasing. In *CVPR Workshops*, 2024.

Colin Lea, Michael D. Flynn, René Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks for action segmentation and detection. In *CVPR*, 2017.

Yiming Li, Xiaoshan Yang, and Changsheng Xu. Dynamic scene graph generation via anticipatory pre-training. In *CVPR*, 2022.

Zhixuan Lin, Evgenii Nikishin, Xu Owen He, and Aaron Courville. Forgetting transformer: Softmax attention with a forget gate. In *ICLR*, 2025.

Vladislav Makarov, Mark Gizetdinov, and Dmitry Yudin. SceneGraphVLM: Dynamic scene graph generation from video with vision-language models. *arXiv preprint arXiv:2605.13667*, 2026.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021.

Sayak Nag, Kyle Min, Subarna Tripathi, and Amit K. Roy-Chowdhury. Unbiased scene graph generation in videos. In *CVPR*, 2023.

Trong-Thuan Nguyen, Pha Nguyen, Jackson Cothren, Alper Yilmaz, and Khoa Luu. HyperGLM: Hypergraph for video scene graph generation and anticipation. In *CVPR*, 2025.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *ACM CHIL*, 2020.

Rohith Peddi, Saurabh, Ayush Abhay Shrivastava, Parag Singla, and Vibhav Gogate. Towards unbiased and robust spatio-temporal scene graph generation and anticipation. In *CVPR*, 2025.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *ICLR*, 2022.

Tao Pu, Tianshui Chen, Hefeng Wu, Yongyi Lu, and Liang Lin. Spatial-temporal knowledge-embedded transformer for video scene graph generation. *IEEE TIP*, 2023.

Raphael Ruschel, Md Awsafur Rahman, Hardik Prajapati, Suya You, and B. S. Manjunath. Temporally consistent dynamic scene graphs: An end-to-end approach for action tracklet generation. *arXiv preprint arXiv:2412.02808*, 2024.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

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

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *NAACL*, 2018.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016.

Judith D. Singer and John B. Willett. It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, 1993.

Shawn Tan, Songlin Yang, Aaron Courville, Rameswar Panda, and Yikang Shen. Scaling stick-breaking attention: An efficient implementation and in-depth study. In *ICLR*, 2025.

Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, 2019.

Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target adaptive context aggregation for video scene graph generation. In *ICCV*, 2021.

Guan Wang, Zhimin Li, Qingchao Chen, and Yang Liu. OED: Towards one-stage end-to-end dynamic scene graph generation. In *CVPR*, 2024.

Shuang Wang, Lianli Gao, Xinyu Lyu, Yuyu Guo, Pengpeng Zeng, and Jingkuan Song. Dynamic scene graph generation via temporal prior inference. In *ACM MM*, 2022.

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.

Shun-Zheng Yu. Hidden semi-Markov models. *Artificial Intelligence*, 2010.