RELATION EVENT MEASURE: PRESERVING EVENT SUPERVISION UNDER PARTIAL GROUNDING FOR PANOPTIC VIDEO SCENE GRAPH GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Panoptic video scene graph generation (PVSG) commonly constructs relation targets by matching annotations to predicted tracklets. When tracklets are incomplete, this matching process can discard entire events or truncate their positive temporal supervision. We introduce **Relation Event Measure (REM)**, a mass-based formulation that decouples the full event annotation from partial visual groundings. To achieve this goal, we build a Relation Event Measure Prediction to separates event count and temporal allocation from grounding availability and tracklet-pair allocation. It preserves supervision for observed mass on matched pairs and unobserved mass in an explicit unobserved state. To supervise relation learning under REM, we propose the Annotation-Preserving Event Learning, which preserves event coverage and annotated temporal extent through measure supervision over all valid events and full-span temporal supervision.Furthermore, REM-based Relation Inference integrates temporal mass to score tracklet-pair relations and uses mass-weighted queries for temporal localization. REM improves strict mR@100 over the Retained REM control by 3.43 and 0.37 percentage points under IPS+T and VPS, respectively.

1 INTRODUCTION

Panoptic video scene graph generation (PVSG) aims to represent a video as a temporally evolving graph, with entities grounded by pixel-level panoptic masks and edges describing their semantic relations over time Yang et al. (2023). PVSG covers both countable objects and background stuff regions as entities and recognizes their predicate with temporal support. The same entity pair may participate in different relations at different times, making temporal relation learning essential to describing how interactions unfold throughout a video.

Existing PVSG pipelines build pairwise relation representations from predicted tracklets and construct supervision by matching them to ground-truth (GT) entities (Yang et al., 2023). MCL improves temporal relation representations through motion-aware contrastive learning and optimal-transport-based tube alignment (Nguyen et al., 2025). SegPVSG introduces temporal focusing and relation-centric generative augmentation, retaining the full temporal annotations of events that survive tracklet matching (Li et al., 2026). Despite these developments, **the availability of relation supervision remains tied to the predicted tracklets through which an event is grounded**.

This dependency conflates two distinct roles: *an annotation defines a complete relation event, whereas predicted tracklets provide observations of that event*. Missed detections, occlusions, segmentation errors, and identity failures can make those observations intermittent or fragmented. Figure 1 illustrates two failures of matching-based construction: Event A retains positive temporal targets only where the subject and object are jointly matched, whereas Event B still has intermittent correspondences but is excluded from training because no candidate pair survives matching and filtering. Preserving full temporal labels for retained events, as in SegPVSG (Li et al., 2026), does not restore events already excluded by this process. The problem is therefore not merely incomplete visual evidence, but the propagation of observation incompleteness into the supervision itself. The event-deletion rates in Figure 1 use the same 3,592 valid training events as the denominator: 440 events (12.2%) are discarded under IPS+T and 1,274 (35.5%) under VPS (Appendix C).

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.

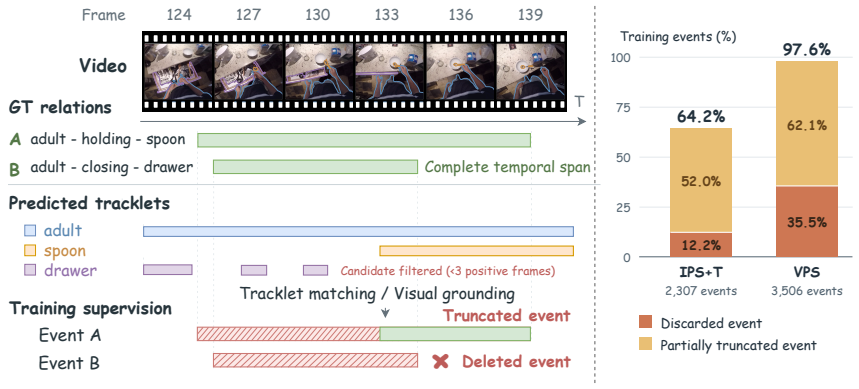

Figure 1: **Event supervision under partial visual grounding.** Matching-based supervision can truncate positive temporal support (A) or discard events (B). Right: a large proportion of training events are discarded or partially truncated under both IPS+T and VPS settings.

To preserve relation supervision under partial visual grounding, we introduce **Relation Event Measure (REM)**, a mass-based formulation that treats each complete annotated relation event as a unit of supervision. Each event contributes one count and one unit of normalized mass over its full annotated temporal span. *Relation Event Measure Prediction* separates event count and temporal allocation from grounding availability and tracklet-pair allocation. Mass is allocated to compatible observed subject–object pairs or an explicit unobserved state, allowing annotations to supervise learning even when their visual grounding is incomplete or unavailable.

To learn from these preserved annotations, *Annotation-Preserving Event Learning* combines measure supervision over all valid events with full-span temporal supervision from available event observations. The measure objective retains event-count, temporal-mass, and grounding-allocation targets regardless of correspondence availability. For events with at least one valid subject–object correspondence, mass-weighted observations form queries for a shared temporal-span predictor trained on the full annotated target. This preserves positive temporal labels even at annotated times without grounding.

REM-based Relation Inference connects this learned measure to relation prediction over the original predicted tracklet pairs. Integrating each pair's assigned mass over its joint observations yields relation scores, while normalized mass-weighted subject and object features form queries for the shared temporal-span predictor. The resulting predictions combine relation ranking with temporal localization while keeping the upstream masks and tracklets fixed.

- We identify how partial tracklet grounding removes available event supervision and introduce REM to preserve each annotated event as normalized temporal mass allocated to observed pairs or an explicit unobserved state.
- We develop annotation-preserving event learning and REM-based inference, connecting measure supervision over all valid events and full-span learning from available observations to relation ranking and temporal localization.
- Experiments on PVSG show that REM improves strict mR@100 over the Retained REM control by 3.43 and 0.37 percentage points under IPS+T and VPS, respectively, supporting the value of preserving event coverage.

2 Related Work

Video relation and scene-graph modeling. Image SGG studies structured context and long-tailed predicates (Zellers et al., 2018; Tang et al., 2019; Lin et al., 2020; Tang et al., 2020; Yu et al., 2021), while video models exploit tracklets, spatiotemporal context, and Transformer-based dynamic graphs (Shang et al., 2017; Liu et al., 2020; Ji et al., 2020; Cong et al., 2021; Li et al., 2022; Zheng et al., 2022; Nag et al., 2023). PVSG adds pixel-level panoptic grounding to temporally localized relations (Yang et al., 2023). MCL improves motion-aware representations (Nguyen et al.,

2025), SegPVSG combines temporal focusing with generative augmentation (Li et al., 2026), and UNO uses one-stage prediction (Le et al., 2026). SegPVSG preserves complete temporal labels for retained relations, but its released preprocessing still excludes events without eligible matched pairs.

Relation supervision under partial observations. Relation learning depends on whether upstream processing preserves sufficient object-pair and temporal evidence. Pair-Net Wang et al. (2024b) identifies pair recall as a bottleneck in panoptic SGG, while PAWS Kang et al. (2026) improves detector-generated pairs in weakly supervised VidSGG. OED Wang et al. (2024a) further highlights the difficulty of aggregating relation evidence across frames. These methods improve observable evidence but still construct supervision from successfully detected or matched entities. Prior SGG work has also addressed limited or biased labels, long-tailed predicates, and weak supervision (Chen et al., 2019; Tang et al., 2020; Chiou et al., 2021; Shi et al., 2021; Chen et al., 2023; Kim et al., 2025). In contrast, we study the loss of *available* annotation information caused by matching-based supervision over incomplete predicted tracklets. Building on the separation between semantic supervision and detector reachability in Dual-SGG (Qu et al., 2026), we extend this principle to temporally localized events by retaining each complete predicate annotation as a supervision unit and assigning its mass to observed pairs or an explicit unobserved state.

3 METHODS

Relation learning for PVSG combines complete annotated relation events with only partially observed predicted tracklets. We propose Relation Event Measure (REM) to preserve each annotated event as a complete supervision unit while accounting for incomplete visual grounding. REM learns from fixed predicted tracklets and complete relation annotations (Figure 2). All valid events contribute to measure supervision. Events with at least one valid subject–object correspondence additionally train a shared temporal-span predictor on their full annotated temporal targets. At inference, integrated mass scores native tracklet pairs, and mass-weighted queries drive the same predictor for temporal localization.

3.1 PROBLEM FORMULATION

Given a video frame set $\mathcal{F}$ and T frames, an event is represented as $e = (u_e, v_e, r_e, S_e)$, where u_e and v_e are the GT subject and object identities, $r_e \in \mathcal{R}$ is the predicate label, and $S_e \subseteq \mathcal{F}$ is its annotated temporal span. $\mathcal{E}$ is the set of valid annotated relation events and $\mathcal{R}$ the predicate category. S_e is defined as the union of the annotated intervals for the same subject–object pair and predicate. The relation-activity indicator $Y_e(t) = \mathbf{1}[t \in S_e]$ specifies whether a relation event e is active at time t.

For each sampled time $t \in \mathcal{T} \subseteq \mathcal{F}$ during relation learning, $\mathcal{A}_t$ is the set of predicted mask observations. Each observation provides a frozen mask-pooled visual feature, predicted category, and mask geometry that remain fixed during relation learning and inference. Tracklet identifiers associate observations across time. The set of eligible subject–object pairs at time t is $\mathcal{Q}_t = \{(i, j) : i, j \in \mathcal{A}_t,\ i \neq j\}$, where i and j index predicted subject and object observations, respectively.

For an annotated event e and sampled time t, $B_{et} \subseteq \mathcal{Q}_t$ indicates the set of predicted subject-object pairs whose observations are matched to the GT identities u_e and v_e, using a mask-IoU threshold of 0.5. $B_{et} = \varnothing$ if $t \notin S_e$. $o_{et} = \mathbf{1}[B_{et} \neq \varnothing]$ is the grounding indicator. GT masks are used only during training to establish these correspondences.

Current matching-based supervision construction couples the relation-activity indicator $Y_e(t)$ and the grounding indicator o_{et} by retaining relations only where valid subject–object correspondences exist, and discards events that do not survive matching. Equivalently, given $\widetilde{Y}_e(t) = Y_e(t)o_{et}$, when $\sum_{t\in\mathcal{T}} \widetilde{Y}_e(t)$ exceeds the matching threshold, event e is retained. Under this construction, $Y_e(t) = 1$ does not imply that the event is visually grounded at sampled time t, i.e., $o_{et} = 1$.

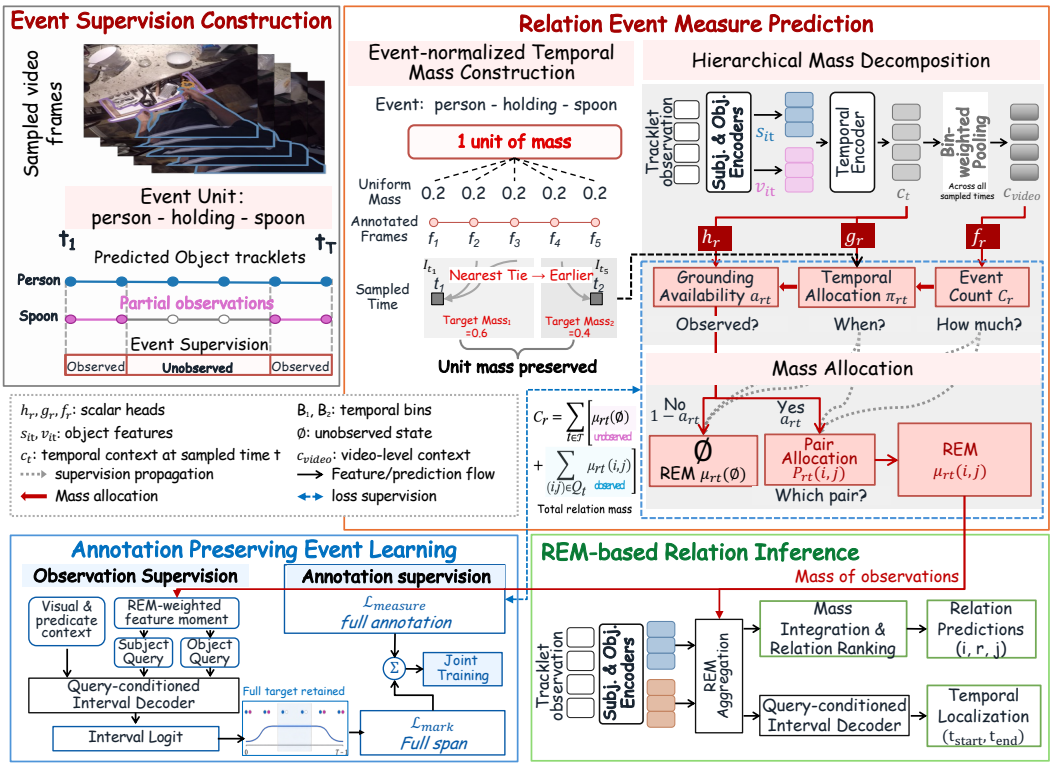

Figure 2: **Overview of Relation Event Measure (REM).** Event annotations define unit-mass temporal targets despite partial tracklet observations. Relation Event Measure Prediction represents each annotated event as unit mass distributed over time, observed pairs, and an unobserved state. In Annotation-Preserving Event Learning, all valid events receive measure supervision, and those with valid correspondences additionally receive full-span temporal supervision. At inference, integrated mass ranks tracklet-pair relations, while mass-weighted queries drive temporal localization.

3.2 Relation Event Measure Prediction

Matching-based supervision may discard annotated events or truncate their temporal supervision when predicated tracklets are incomplete. We introduce Relation Event Measures (REM), which assigns each annotated event one unit of temporal mass, allocated across observed subject–object pairs and an explicit unobserved state.

Event-normalized temporal mass construction.

To preserve annotation-defined supervision while adapting its grounding to available observations, relation-event mass is defined as a nonnegative measure that separates total event mass from its allocation and is additive over disjoint observations. Since the relation model operates on sampled times while relation annotations are defined over the original video frames, we first map the video frame set $\mathcal{F}$ onto the sampled frame set $\mathcal{T}$ by assigning frames to the nearest sampled time, with ties resolved toward the earlier sample. The disjoint temporal bin at sampled time $t \in \mathcal{T}$ is defined as:

$$I_t = \left\{ f \in \mathcal{F} : t = \min\left(\arg\min_{s \in \mathcal{T}} |f - s| \right) \right\}, \qquad b_t = |I_t|, \tag{1}$$

where f indexes an original video frame and b_t is the number of frames represented by sampled time t, with $\sum_{t \in \mathcal{T}} b_t = T$. For each event e, its annotated temporal support S_e is converted to normalized temporal mass w_{et} over sampled times:

$$w_{et} = \frac{|S_e \cap I_t|}{|S_e|}, \qquad \sum_{t \in \mathcal{T}} w_{et} = 1. \tag{2}$$

Therefore, an event annotation contributes exactly one unit of temporal supervision, regardless of its duration or the number of tracklet fragments through which it is observed. Compared to the sampled binary activity label $Y_e(t)$, the normalized mass w_{et} accounts for all annotated frames, preserving their contribution to temporal-mass supervision even when short intervals fall between sampled frames.

Hierarchical mass decomposition. Directly assigning mass to observations would couple the existence and temporal distribution with grounding availability again. When no valid pair exists, such a direct assignment provides no support on which to place the event mass. To this end, we incorporate a hierarchical mass decomposition in REM, which factorizes the event mass along a conditional allocation chain, progressing from temporal allocation, grounding availability to pairwise assignment.

Given an observation i at time t, $\boldsymbol{s}_{it}$ and $\boldsymbol{v}_{it}$ denote its subject and object features processed by subject and object encoders, respectively. At each sampled time t, we average $(\boldsymbol{s}_{it} + \boldsymbol{v}_{it})/2$ over all visible objects i to obtain a frame-level representation. The resulting sequence is passed through layer normalization and a temporal Transformer with fixed absolute time encoding, yielding temporal context $\boldsymbol{c}_t$. Empty object sets contribute a zero mean and the video-level context mean is $\boldsymbol{c}_{\text{video}} = \frac{1}{T}\sum_{t\in\mathcal{T}} b_t \boldsymbol{c}_t$.

Three predicate-specific scalar heads f_r, g_r, and h_r parameterize the decomposition. The three factors are defined as:

$$C_r = \exp f_r(\boldsymbol{c}_{\text{video}}), \quad \pi_{rt} = \frac{b_t \exp g_r(\boldsymbol{c}_t)}{\sum_{s\in\mathcal{T}} b_s \exp g_r(\boldsymbol{c}_s)}, \quad a_{rt} = \mathbf{1}[\mathcal{Q}_t \neq \varnothing]\, \sigma(h_r(\boldsymbol{c}_t)), \tag{3}$$

where $C_r > 0$ predicts the number of predicate-r events and determines their total mass, π_{rt} distributes this mass along represented video time, and $a_{rt} \in [0, 1]$ specifies the fraction that can be assigned to visible subject-object pairs. The indicator enforces $a_{rt} = 0$ when no eligible pair is available.

Mass allocation. Given these three factors, REM allocates the observed mass among eligible subject–object pairs. For a nonempty pair set $\mathcal{Q}_t$ at sampled time t, predicate-specific projections ϕ_r^s, ϕ_r^v are constructed to project the context-conditioned subject and object features into a shared pair-key space with dimension $d_p = 8$. Then pairwise allocation probability P_{rt} is computed based on the scaled dot-product similarity ℓ_{rtij}, written as:

$$P_{rt}(i,j) = \frac{\exp \ell_{rtij}}{\sum_{(k,l)\in\mathcal{Q}_t} \exp \ell_{rtkl}}, \qquad \ell_{rtij} = \frac{\phi_r^s(\boldsymbol{s}_{it} + \boldsymbol{c}_t)^\top \phi_r^v(\boldsymbol{v}_{jt} + \boldsymbol{c}_t)}{\sqrt{d_p}}. \tag{4}$$

When $\mathcal{Q}_t$ is empty, $P_{rt}(i,j) = 0$. The predicted event measure over observed pairs and the unobserved state is:

$$\mu_{rt}(i,j) = C_r \pi_{rt} a_{rt} P_{rt}(i,j), \qquad \mu_{rt}(\varnothing) = C_r \pi_{rt}(1 - a_{rt}), \tag{5}$$

where $\mu_{rt}(i,j)$ denotes the mass assigned to an observed subject-object pair and $\mu_{rt}(\varnothing)$ denotes the mass assigned to the unobserved state. $\sum_{t\in\mathcal{T}} \left[\sum_{(i,j)\in\mathcal{Q}_t} \mu_{rt}(i,j) + \mu_{rt}(\varnothing)\right] = C_r$. The unobserved state therefore participates in event learning even when no valid correspondence is available.

3.3 Annotation-Preserving Event Learning

Annotation-preserving supervision. Let $n_r = |\{e \in \mathcal{E} : r_e = r\}|$ be the number of events with predicate annotation r in a video and N_r be their count over the training set. The predicate category weight is $\alpha_r = \max(N_r, 1)^{-1/2}$ and $Z = \sum_r \alpha_r$. When a predicate-indexed prediction is evaluated for event e, we set $r = r_e$ thereby obtaining $P_{r_e,t}(i,j) \equiv P_{r,t}(i,j)|_{r=r_e}$. Thus, $P_{r_e,t}$ is shared by events with the same predicate rather than predicted separately for each event. For an event e at sampled time t, its grounding likelihood is:

$$\gamma_{et} = \begin{cases} \log a_{r_e t} + \log \displaystyle\sum_{(i,j)\in B_{et}} P_{r_e t}(i,j), & o_{et} = 1, \\ \log(1 - a_{r_e t}), & o_{et} = 0. \end{cases} \tag{6}$$

When correspondences are available, the event learning is supervised by the total probability mass assigned to all compatible subject–object pairs. Otherwise, it is supervised with the mass assigned to the unobserved state. The resulting measure objective is

$$\mathcal{L}_{\text{measure}} = \frac{1}{Z}\sum_{r}\alpha_r\left(C_r - n_r \log C_r\right) - \frac{1}{Z}\sum_{e}\alpha_{r_e}\sum_{t\in\mathcal{T}} w_{et}\left[\log \pi_{r_e t} + \gamma_{et}\right]. \tag{7}$$

Under this formulation, correspondence availability affects only how an event is grounded; events without any valid subject–object correspondence still supervise event count, temporal mass, and the unobserved state.

Full-span observation supervision. For each event with at least one valid correspondence, we aggregate its available mass-weighted evidence into an event-level query and train a shared temporal-span predictor on the complete annotated target, retaining positive labels at annotated positive times even without grounding. Given the set B of observed pairs, the total assigned mass $M_r(B)$, feature moments $\boldsymbol{U}_r(B)$ and $\boldsymbol{V}_r(B)$ of subject and object are defined as:

$$M_r(B) = \sum_{(t,i,j)\in B}\mu_{rt}(i,j), \quad \boldsymbol{U}_r(B) = \sum_{(t,i,j)\in B}\mu_{rt}(i,j)\boldsymbol{s}_{it}, \quad \boldsymbol{V}_r(B) = \sum_{(t,i,j)\in B}\mu_{rt}(i,j)\boldsymbol{v}_{jt}, \tag{8}$$

When the set of observations of event e is non-empty, i.e., $B_e \neq \varnothing$, mass-normalized subject and object queries are constructed as:

$$\boldsymbol{q}_e^s = \frac{\boldsymbol{U}_{r_e}(B_e)}{M_{r_e}(B_e)}, \qquad \boldsymbol{q}_e^o = \frac{\boldsymbol{V}_{r_e}(B_e)}{M_{r_e}(B_e)}. \tag{9}$$

These queries summarize the available visual evidence for event e according to the learned event measure, regardless of whether that evidence is contained in a single tracklet pair or distributed across multiple fragments. The shared temporal-span predictor, implemented as a query-conditioned interval decoder, maps these queries to interval logits $z_e(t)$ over the full sampled video. It retrieves subject and object evidence from each visible object set and integrates this evidence with the queries, temporal context, and predicate embedding. Appendix B.1 details the decoding process.

For each event e, let $n_e^+ = \sum_{t\in\mathcal{T}} Y_e(t)$ and $n_e^- = \sum_{t\in\mathcal{T}}(1 - Y_e(t))$ denote the numbers of positive and negative sampled times. The full annotated temporal target is supervised with a loss that balances these two groups:

$$\mathcal{L}_{\text{mark}} = -\frac{1}{Z}\sum_{e:B_e\neq\varnothing}\alpha_{r_e}\left[\frac{\sum_{t:Y_e(t)=1}\log\sigma(z_e(t))}{\max(1,n_e^+)} + \frac{\sum_{t:Y_e(t)=0}\log\sigma(-z_e(t))}{\max(1,n_e^-)}\right]. \tag{10}$$

Events with no valid correspondence contribute through $\mathcal{L}_{\text{measure}}$, while events with at least one valid correspondence additionally supervise the temporal-span predictor. All relation parameters are optimized jointly with $\mathcal{L} = \mathcal{L}_{\text{measure}} + \mathcal{L}_{\text{mark}}$.

3.4 REM-BASED RELATION INFERENCE

At inference, GT object correspondences for annotated relation events are unavailable. REM therefore returns to the native tracklet space, i.e., the unmodified tracklet identities. For each predicted tracklet pair (i, j), its relation score is obtained by integrating the learned event measure over all of its co-observations. Let $B_{ij} = \{(t,i,j) : t \in \mathcal{T},\ (i,j) \in \mathcal{Q}_t\}$ denote the set of sampled times at which the pair is jointly observed. The predicate-r relation score is:

$$S_{ij\to r} = \alpha_r M_r(B_{ij}) = \alpha_r C_r \sum_{t\in\mathcal{T}}\pi_{rt}a_{rt}P_{rt}(i,j). \tag{11}$$

We retain at most 100 tracklet pairs ranked by their maximum predicate scores, and further rank the resulting subject–object–predicate predictions by $S_{ij\to r}$. Temporal localization reuses the temporal-span predictor learned during training. For each prediction, we construct subject and object queries as $\boldsymbol{U}_r(B_{ij})/M_r(B_{ij})$ and $\boldsymbol{V}_r(B_{ij})/M_r(B_{ij})$, and feed them into the shared temporal-span predictor to produce $z_{ijr}(t)$. The resulting logits are interpolated to the original video frame rate to obtain the final temporal span.

Table 1: Comparison on the PVSG dataset; the best are shown in bold. † denotes the published SegPVSG variant without RGVA.

	vIoU ≥ 0.5			vIoU ≥ 0.1		
Method	R/mR@20	R/mR@50	R/mR@100	R/mR@20	R/mR@50	R/mR@100
One-stage:						
UNO ResNet-50	6.23/5.60	7.37/6.84	8.65/8.21	13.83/11.65	19.27/15.94	24.63/19.99
IPS+T:						
Transformer	3.88/2.81	5.66/4.12	6.18/4.44	9.01/6.69	14.88/11.28	17.51/13.20
Convolution	3.88/2.55	5.24/3.29	6.71/5.36	10.06/8.98	14.99/12.21	18.13/15.47
MCL Transformer	3.98/2.98	5.97/4.20	7.44/5.15	10.59/9.56	16.98/12.39	22.33/17.47
MCL Convolution	4.51/3.56	6.08/4.38	7.76/5.86	11.43/9.57	17.30/13.13	22.85/17.48
SegPVSG†	5.14/6.67	9.12/10.15	12.58/12.87	12.16/13.20	21.38/21.58	27.57/26.45
REM (Ours)	**8.39/7.52**	**12.05/12.75**	**13.63/14.28**	**18.55/15.59**	**26.42/22.89**	**30.29/26.78**
VPS:						
Transformer	0.42/0.61	0.73/0.76	1.05/0.92	6.50/5.75	9.64/8.25	12.26/9.51
Convolution	0.42/0.25	0.63/0.67	0.63/0.67	8.07/7.84	11.01/9.78	12.89/10.77
MCL Transformer	0.63/0.83	1.05/0.76	1.05/0.76	6.71/6.94	10.27/8.68	13.42/12.09
MCL Convolution	0.84/0.98	1.26/1.22	1.26/1.22	8.18/8.00	12.90/11.47	14.22/13.59
SegPVSG†	1.78/1.52	2.52/2.16	3.04/2.79	8.39/8.45	15.72/15.33	22.01/21.63
REM (Ours)	**2.10/1.65**	**2.62/2.24**	**3.14/2.81**	**14.47/12.57**	**19.92/17.76**	**24.00/23.35**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Dataset, implementation, and evaluation. We evaluate on PVSG (Yang et al., 2023), comprising 338 training and 62 validation videos from VidOR (Shang et al., 2019), EPIC-KITCHENS (Damen et al., 2018), and Ego4D (Grauman et al., 2022). Following the standard IPS+T and VPS settings (Yang et al., 2023), we use the released masks, categories, tracklet identities, and Mask2Former ResNet-50 features (Cheng et al., 2022). We report PVSG R@K and mR@K at vIoU thresholds of 0.5 and 0.1, together with temporal-gated G-R/G-mR requiring tIoU ≥ 0.5 with the full annotated relation span, and mean tIoU over matched candidate–GT tuples to evaluate temporal localization independently of top-K recall.

Compared methods. We compare REM with the original PVSG baselines (Yang et al., 2023), MCL (Nguyen et al., 2025), UNO (Le et al., 2026) and SegPVSG (Li et al., 2026). For controlled analysis, we additionally use three local baselines: *Retained REM*, which keeps the REM architecture but trains only on events retained from grounding with the annotated relations; *Simple BCE*, which uses all valid events and their full annotated relation spans with the same temporal-span readout, but replaces event-mass allocation with pair–predicate BCE; and *Strong Conv*, which uses supervision produced by tracklet grounding together with strengthened frozen inputs and local scoring. Appendix B.2 provides specific implementation details.

4.2 COMPARISON WITH PRIOR WORK

Table 1 reports published PVSG comparisons at $K = 20/50/100$. With IPS+T, REM exceeds the original PVSG, MCL, and SegPVSG† results at vIoU ≥ 0.5. At $K = 100$, REM reaches 13.63/14.28 R/mR, compared with 12.58/12.87 for SegPVSG†. With VPS, REM reaches 3.14/2.81 R/mR@100 at vIoU ≥ 0.5, compared with 3.04/2.79 for SegPVSG†. This demonstrates the rationale for fully using annotated relation supervision while also highlighting the outstanding performance of our method. To ensure consistency in the training data, the main table lists the results for SegPVSG† (i.e., the published variant without RGVA), while the full SegPVSG results incorporating RGVA augmentation are presented separately in Appendix E.2.

Table 2: Comparison of supervision interfaces under vIoU ≥ 0.5.

	Model	R/mR@20	R/mR@50	R/mR@100	G-mR@100	tIoU
IPS+T	REM	8.39/7.52	12.05/12.75	13.63/14.28	6.75	0.373
	Retained REM	8.91/6.86	11.64/9.55	12.68/10.85	4.23	0.361
	Simple BCE (top-10%)	1.89/4.83	4.82/7.67	8.39/9.54	4.79	0.369
	Strong Conv	6.08/4.84	7.55/5.55	9.12/6.59	4.71	–
VPS	REM	2.10/1.65	2.62/2.24	3.14/2.81	0.56	0.232
	Retained REM	1.78/1.15	2.62/2.12	2.94/2.44	0.48	0.208
	Simple BCE (top-10%)	0.31/0.86	0.52/1.14	1.05/1.90	0.00	0.227
	Strong Conv	1.57/1.41	2.20/2.09	2.62/2.47	0.17	–

Table 3: Event-retention, query, and scoring ablations at vIoU ≥ 0.5.

	IPS+T		VPS	
Variant	R/mR@100	G-mR@100	R/mR@100	G-mR@100
Supervision and query construction				
Retained REM	12.68 / 10.85	4.23	2.94 / 2.44	0.48
+ deleted events with sampled correspondences	13.10 / 12.61	4.78	2.73 / 2.55	0.48
+ deleted events without sampled correspondences	13.42 / 11.89	4.78	3.04 / 2.84	0.61
REM (all valid events)	13.63 / 14.28	6.75	3.14 / 2.81	0.56
REM with uniform interval queries	13.10 / 13.84	6.67	2.94 / 3.08	0.54
Relation scoring				
Simple BCE with mean pooling	8.18 / 8.08	3.83	0.94 / 1.03	0.05
Simple BCE with max pooling	8.07 / 9.77	4.68	1.15 / 1.98	0.00
Simple BCE with top-10% pooling	8.39 / 9.54	4.79	1.05 / 1.90	0.00
REM without video-level count C_r	5.24 / 10.25	5.77	1.15 / 1.58	0.37

4.3 Controlled Comparison of Supervision Interfaces

Table 2 tests whether deleted annotations remain useful under fixed observations, and whether retaining full annotated labels suffices without an event-level formulation. Retained REM preserves the full annotated span of each retained event and differs from REM only in the supervised event set. Restoring deleted events raises mR/G-mR@100 from 10.85/4.23 to 14.28/6.75 with IPS+T and from 2.44/0.48 to 2.81/0.56 with VPS, supporting the benefit of decoupling event supervision from visual grounding. The smaller VPS gain is consistent with its lower fixed-tracklet spatial recoverability (Appendix C.2).

Simple BCE is a complementary control: it uses all valid events, complete temporal targets, and auxiliary event-count and temporal-mass supervision, but learns local pair–predicate probabilities and pools them for ranking. REM improves mR@100 over Simple BCE by 4.74 points with IPS+T and 0.91 points with VPS, while mean tIoU on common spatially eligible candidates remains similar. Hence, the advantage is not explained by a uniformly better temporal decoder, motivating the event-measure scoring analysis below. Strong Conv is an architecture reference rather than a supervision-matched control because its losses and class-count conventions differ. Appendix E provides common-candidate, fixed-score, and qualitative diagnostics.

4.4 Ablation Studies

Ablations examine which deleted events remain useful, whether their benefit requires measure-weighted interval queries, and how event evidence enters relation scoring. In Table 3, the two "+ deleted" variants independently add events with or without valid sampled correspondences to Retained REM. Supervision and query variants are trained separately under the same budget within each setting, while scoring controls reuse frozen models.

Event coverage and interval queries. Events without sampled correspondences test whether annotations must be visually grounded to remain useful. Adding them raises mR@100 from 10.85 to

Table 4: Relation-scoring controls under Simple BCE. C_r denotes video-level event count, π learned temporal allocation, and π^{unif} the uniform-allocation control.

	Variant	R/mR@20	R/mR@50	R/mR@100	G-mR@100
IPS+T	REM	8.39/7.52	12.05/12.75	13.63/14.28	6.75
	Simple BCE (top-10%)	1.89/4.83	4.82/7.67	8.39/9.54	4.79
	Simple BCE + C_r	3.56/3.30	5.35/4.44	7.86/6.32	3.20
	Simple BCE + π^{unif}	7.86/7.45	11.01/9.43	13.10/10.30	5.26
	Simple BCE + π	6.71/7.38	10.48/9.29	12.58/11.25	5.20
	Simple BCE + C_r + π	6.92/5.98	9.96/8.04	11.95/9.98	4.66
VPS	REM	2.10/1.65	2.62/2.24	3.14/2.81	0.56
	Simple BCE (top-10%)	0.31/0.86	0.52/1.14	1.05/1.90	0.00
	Simple BCE + C_r	0.84/0.69	1.26/0.98	1.47/1.00	0.13
	Simple BCE + π^{unif}	1.89/1.65	2.52/2.11	2.73/2.34	0.17
	Simple BCE + π	1.89/1.65	2.62/2.16	2.94/2.45	0.17
	Simple BCE + C_r + π	1.89/1.20	2.31/1.58	2.73/2.21	0.17

11.89 with IPS+T and from 2.44 to 2.84 with VPS, despite being unable to form interval queries. Their gain therefore comes from event-count, temporal-mass, and unobserved-state supervision rather than additional grounded interval targets. Adding deleted events with sampled correspondences yields 12.61 and 2.55 mR@100, respectively, and additionally supervises the full temporal span from available observations. Both results show that tracklet-level label grounding discards useful supervision.

With IPS+T, combining both event types performs best at 14.28 mR and 6.75 G-mR. VPS is non-monotonic: all-event REM reaches 2.81/0.56, slightly below the no-sampled-correspondence variant at 2.84/0.61. Uniform interval queries also underperform REM with IPS+T, but improve VPS mR to 3.08 while slightly reducing G-mR to 0.54. The benefit of retaining full event annotations therefore does not require measure-weighted queries, and the results do not establish that such weighting is uniformly superior.

Relation scoring. Preserving full events during training should also be reflected in how relation hypotheses are ranked. Mean pooling can dilute a short relation's responses across a longer co-visible tracklet span; max and top-10% pooling reduce this averaging effect but remain below REM in both settings. Within REM, removing only the video-level count factor C_r lowers mR/G-mR from 14.28/6.75 to 10.25/5.77 with IPS+T and from 2.81/0.56 to 1.58/0.37 with VPS. This supports using video-level predicate evidence together with temporal and pair allocation, as in Equation 11.

Event-level outputs versus joint event-mass learning. Table 4 tests whether incorporating event-level outputs into BCE scoring accounts for REM's advantage. Learned temporal allocation improves Simple BCE mR@100 from 9.54 to 11.25 with IPS+T and from 1.90 to 2.45 with VPS, but remains below REM at 14.28 and 2.81. Adding C_r alone reduces mR to 6.32 and 1.00, while combining C_r and π reaches 9.98 and 2.21. Thus, event-count information is not a universally beneficial score correction: it helps within the learned factorization of REM but does not provide the same benefit when added to BCE scores. Together with the preceding ablations, these results support linking full event supervision to the joint allocation of event mass, rather than treating count prediction or temporal weighting as interchangeable additions to a local classifier.

5 Conclusion

We introduced Relation Event Measure (REM), a mass-based formulation for preserving complete-event supervision under partial visual grounding in PVSG. By allocating event mass to observed pairs or an explicit unobserved state, REM retains measure supervision for all valid events and supervises full-span prediction for those with available correspondences. The same measure supports relation ranking through mass integration and temporal localization through mass-weighted queries on predicted tracklet pairs. Results under IPS+T and VPS support the value of separating annotation-defined supervision from grounding availability while keeping upstream masks and tracklets fixed.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and improve the clarity and presentation of the manuscript. All AI-assisted content was reviewed and revised by the authors, who take full responsibility for the final text.

REPRODUCIBILITY STATEMENT

The source bundle contains code, configurations, per-video evidence, table and figure generators, and LaTeX sources. Its reproduction guide links public OpenPVSG data/model entry points and provides commands for tracking, feature/target generation, and relation training with a configurable data root. Observation-cache checksums, the implementation, and evaluation are documented in Appendices B-E; recoverability and sampling-sensitivity analyses are also included.

BROADER IMPACT

This study investigates supervision construction for structured video understanding using public PVSG data and does not collect new human-subject data. Relation predictions inherit upstream segmentation/tracking errors, limited dataset coverage, and social biases. Benchmark accuracy should therefore not be interpreted as reliable evidence about a person's intentions or behavior.

REFERENCES

Siqi Chen, Jun Xiao, and Long Chen. Video scene graph generation from single-frame weak supervision. In *International Conference on Learning Representations*, 2023.

Vincent S Chen, Paroma Varma, Ranjay Krishna, Michael Bernstein, Christopher Re, and Li Fei-Fei. Scene graph prediction with limited labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2580–2590, 2019.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1280–1289. IEEE, 2022.

Meng-Jiun Chiou, Henghui Ding, Hanshu Yan, Changhu Wang, Roger Zimmermann, and Jiashi Feng. Recovering the unbiased scene graphs from the biased ones. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 1581–1590, 2021.

Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16352–16362. IEEE, 2021.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: the dataset. In *European conference on computer vision*, pp. 753–771. Springer, 2018.

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*, pp. 18995–19012, 2022.

Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10233–10244. IEEE, 2020.

Minseok Kang, Minhyeok Lee, Minjung Kim, Jungho Lee, Donghyeong Kim, Sungmin Woo, Inseok Jeon, and Sangyoun Lee. Revisiting weakly-supervised video scene graph generation via pair affinity learning. In *European Conference on Computer Vision*, 2026.

Kibum Kim, Kanghoon Yoon, Yeonjun In, Jaehyeong Jeon, Jinyoung Moon, Donghyun Kim, and Chanyoung Park. Weakly supervised video scene graph generation via natural language supervision. In *International Conference on Learning Representations*, volume 2025, pp. 15480–15502, 2025.

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 *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2797–2807. IEEE, 2026.

YiKai Li, Quhui Ke, Jinglin Liang, Zhi-Yuan Zhang, Zhidi Lin, and Shuangping Huang. Segpvsg: Panoptic video scene graph generation via temporal focusing and generative augmentation. In *Forty-third International Conference on Machine Learning*, 2026.

Yiming Li, Xiaoshan Yang, and Changsheng Xu. Dynamic scene graph generation via anticipatory pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13864–13873. IEEE, 2022.

Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3743–3752. IEEE, 2020.

Chenchen Liu, Yang Jin, Kehan Xu, Guoqiang Gong, and Yadong Mu. Beyond short-term snippet: Video relation detection with spatio-temporal global context. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10837–10846. IEEE, 2020.

Sayak Nag, Kyle Min, Subarna Tripathi, and Amit K Roy-Chowdhury. Unbiased scene graph generation in videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22803–22813. IEEE, 2023.

Thong Thanh Nguyen, Xiaobao Wu, Yi Bin, Cong-Duy T Nguyen, See-Kiong Ng, and Anh Tuan Luu. Motion-aware contrastive learning for temporal panoptic scene graph generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6218–6226, 2025.

Runfeng Qu, Pia K Bideau, Ole Hall, Julie Ouerfelli-Ethier, Klaus Obermayer, and Olaf Hellwich. Revisiting scene graph generation from the perspective of detector-conditioned reachability. *arXiv preprint arXiv:2607.06176*, 2026.

Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1300–1308, 2017.

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

Jing Shi, Yiwu Zhong, Ning Xu, Yin Li, and Chenliang Xu. A simple baseline for weakly-supervised scene graph generation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16373–16382. IEEE, 2021.

Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6612–6621. IEEE, 2019.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3713–3722. IEEE, 2020.

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 (CVPR)*, pp. 27938–27947, June 2024a.

Jinghao Wang, Zhengyu Wen, Xiangtai Li, Zujin Guo, Jingkang Yang, and Ziwei Liu. Pair then relation: Pair-net for panoptic scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10452–10465, 2024b.

Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, and Luca Bertinetto. Do different tracking tasks require different appearance models? In *Advances in Neural Information Processing Systems*, pp. 726–738, 2021.

Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, et al. Panoptic video scene graph generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18675–18685. IEEE, 2023.

Jing Yu, Yuan Chai, Yujing Wang, Yue Hu, and Qi Wu. Cogtree: Cognition tree loss for unbiased scene graph generation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 1274–1280, 2021.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 5831–5840. IEEE, 2018.

Sipeng Zheng, Shizhe Chen, and Qin Jin. Vrdformer: End-to-end video visual relation detection with transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18814–18824. IEEE, 2022.

APPENDIX

A THEORETICAL PROPERTIES

A.1 INTERPRETATION OF THE MEASURE SURROGATE

Let $A_e \in \{0, 1\}$ indicate event retention by tracklet-level grounding, and reuse the grounding indicator o_{et} from Section 3. Deleting events with $A_e = 0$ removes their supervision; encoding unavailable positives as zeros instead yields $\widetilde{Y}_e(t) = Y_e(t)A_e o_{et}$. Writing X for the model input and suppressing event/time indices, the probability chain rule gives

$$\Pr(\widetilde{Y} = 1 \mid X) = \Pr(Y = 1 \mid X) \Pr(Ao = 1 \mid Y = 1, X). \tag{12}$$

Thus, this target represents annotated activity together with retained grounding availability. This argument applies only to zero-filled targets; it does not equate event deletion or ignored labels with negative supervision.

REM retains every valid event and separates activity $Y_e(t)$, normalized temporal mass w_{et}, and grounding availability o_{et}. The count term in Equation 7 is a Poisson negative log-likelihood up to the constant $\log(n_r!)$. Under w_{et}, the temporal and grounding terms are expected negative log scores for a sampled bin and its valid correspondence set or unobserved state. Each event contributes one normalized supervision unit, regardless of its duration or tracklet fragmentation. This interpretation does not imply a calibrated joint probability model over complete events.

For the measure loss alone, define total temporal mass $W_{rt} = \sum_{e:r_e=r} w_{et}$ and observed temporal mass $O_{rt} = \sum_{e:r_e=r} w_{et}o_{et}$. With unrestricted outputs and $n_r > 0$, the optima satisfy

$$\begin{aligned} C_r^* &= n_r, & \pi_{rt}^* &= \frac{W_{rt}}{n_r}, \\ a_{rt}^* &= \frac{O_{rt}}{W_{rt}}, & C_r^* \pi_{rt}^* a_{rt}^* &= O_{rt}, \end{aligned} \tag{13}$$

for $W_{rt} > 0$, with boundary optima understood as limits. These identities need not hold for the jointly trained neural model. Here, a_{rt} represents the fraction of predicate-r mass allocated to observed pairs, not an entity-specific missingness probability.

Events with no valid correspondence still supervise counts, temporal allocation, and the unobserved state, without assigning unsupported spatial identities. Event mass is not a binary activity label: absent observed mass can reflect either inactivity or unavailable grounding. Accordingly, the interval readout is supervised only for events with at least one valid correspondence, using their full annotated relation spans.

A.2 FIXED-OBSERVATION PARTITION CONSISTENCY

The supervision principle in Section 3 implies that reindexing unchanged observations should not change the contribution of an annotated event to learning. We formalize this property by splitting a predicted tracklet into disjoint temporal fragments while keeping its masks, features, categories, geometry, sampled times, and local GT correspondences fixed. Only the tracklet identifiers assigned to the observations change.

Proposition 1 (Fixed-observation partition consistency) *Suppose that the per-frame observations, features, predicted categories, geometry, sampled times, annotations, and local GT correspondences are fixed, and that each tracklet is replaced by disjoint temporal fragments that exactly partition its observations. With stochastic layers disabled, the objective defined by Equations 7 and 10, together with its parameter and input gradients wherever differentiable, remains unchanged. Event mass and feature moments are additive across the fragments, so recombining their statistics recovers the original event representation and interval logits.*

Proof. Let ρ map each fragment to its original tracklet identifier. A disjoint temporal partition preserves a one-to-one correspondence between visible observations at each time and does not add

or remove eligible pairs. The shared set encoders are equivariant to this relabeling. Hence the visible-set mean, temporal context, C_r, π_{rt}, a_{rt}, pair logits, P_{rt}, and corresponding event-mass values μ_{rt} are unchanged after mapping indices through ρ. The targets $\mathcal{E}$, n_r, w_{et}, and $Y_e(t)$ are also unchanged, and valid correspondence sets contain the same observations under the new indices. The probability sums in Equation 7 are therefore identical.

For an original pair (i, j), let $\mathcal{F}_{ij}$ contain the fragment pairs that map to it, and let B'_{ab} be the observation set of fragment pair (a, b). The additivity in Equation 8 gives

$$M_r(B_{ij}) = \sum_{(a,b)\in\mathcal{F}_{ij}} M'_r(B'_{ab}), \quad U_r(B_{ij}) = \sum_{(a,b)\in\mathcal{F}_{ij}} U'_r(B'_{ab}), \quad V_r(B_{ij}) = \sum_{(a,b)\in\mathcal{F}_{ij}} V'_r(B'_{ab}). \tag{14}$$

The same additivity holds for each event's correspondence set B_e. Aggregating mass and feature moments before normalization therefore recovers the original event queries and interval logits, leaving Equation 10 unchanged. Because the objectives are identical functions of the fixed observations, their parameter gradients and correspondingly indexed input gradients are also identical wherever differentiable.

Scope and numerical check. The property concerns reindexing, not changes to visual evidence, and does not imply identical rankings of fragments left unrecombined. Inference uses native tracklet pairs without GT-dependent merging. With float64 and dropout disabled, splitting each tracklet into two fragments on three training videos per setting gives maximum loss differences of 5.55×10^{-17} (IPS+T) and zero (VPS). Relative parameter/input-gradient differences are $1.53 \times 10^{-15}/3.40 \times 10^{-15}$ and $1.25 \times 10^{-15}/2.08 \times 10^{-15}$, respectively; event-mass differences remain below 10^{-15}.

B Implementation and Baselines

Observations and targets. IPS+T uses the released IPS masks and tracks (Wang et al., 2021) with Mask2Former ResNet-50 pixel-decoder weights (Cheng et al., 2022); VPS uses the supplied video masks, tracks, and corresponding weights. We uniformly sample at most 128 integer frames including both endpoints, resize masks to the feature grid, and average the 256-dimensional pixel-decoder features within each mask. Category embeddings and ten normalized mask-geometry values are projected to a 256-dimensional residual. All controlled methods share these frozen inputs, without GT-mask features, synthetic appearances, tracklet completion, or duplicated masks. The VPS targets are reconstructed exactly for all 338 training videos; raw-mask matching and feature extraction also pass spot checks.

Intervals sharing an ordered GT entity pair and predicate are unioned under the benchmark's integer-frame convention. Removing 24 annotations outside the 57-predicate vocabulary and two events with empty in-video support leaves 3,592 valid events. For 32 events, no sampled frame lies inside the annotated support, but Equation 2 preserves their temporal mass in the unobserved state. Predicted-category agreement is not required for training correspondences. Write $B_e = \{(t, i, j) : (i, j) \in B_{et}\}$; only events with $B_e \neq \varnothing$ contribute interval loss.

Network and numerical conventions. Separate subject/object encoders are two-layer Transformers with width 256, eight heads, and feed-forward width 512. A two-layer temporal Transformer with the same dimensions processes visible-set means. Sinusoidal absolute-time encoding uses $t/\max(1, T-1)$. Predicate-specific pair keys have dimension eight. The complete relation network has 4,010,108 trainable parameters. Count logits are clipped to $[-12, 8]$ before exponentiation. Pair normalization excludes ineligible pairs; empty pair sets route all mass to the unobserved state. The common count factor cancels analytically in normalized training queries; correspondence-aggregation and normalization denominators are floored at 10^{-12}.

Optimization. Relation parameters are randomly initialized, with predicate log-count biases set from mean training-event counts per video and gate biases set to $\log 4$. Retained-event controls use the same initialization rule and fixed class weights. Adam uses learning rate 10^{-4}, default betas, no weight decay, and no learning-rate schedule. Videos are reshuffled, with gradient accumulation over 32 videos and actual-size normalization for the last partial group. Each run has 1,100 updates and uses all sampled times and predicted objects. Main REM, Retained REM, and Simple BCE

results average three runs using the same seeds within each observation setting. Experiments use a V100-32G, PyTorch 1.11.0+cu113, and CUDA 11.3.

B.1 Query-conditioned interval decoder

Learned affine key/value projections $K_s, D_s, K_o, D_o : \mathbb{R}^{256} \to \mathbb{R}^{64}$ have separate subject/object parameters; queries and observations share keys within each role. For the subject role and a nonempty visible set,

$$\begin{aligned} \widehat{q}^s_e &= \mathrm{norm}(K_s(\boldsymbol{q}^s_e)), \qquad \widehat{k}^s_{kt} = \mathrm{norm}(K_s(\boldsymbol{s}_{kt})), \\ \omega^s_{etk} &= \mathrm{softmax}_{k\in\mathcal{A}_t}\left(8(\widehat{q}^s_e)^\top \widehat{k}^s_{kt}\right), \end{aligned} \tag{15}$$

where $\mathrm{norm}(x) = x/\max(\|x\|_2, 10^{-12})$. Retrieval gives $\widetilde{s}_{et} = \sum_{k\in\mathcal{A}_t} \omega^s_{etk} D_s(\boldsymbol{s}_{kt})$; the object role analogously yields $\widehat{q}^o_e$ and $\widetilde{v}_{et}$. Retrieved vectors are zero for an empty visible set. At each time, concatenate six 64-dimensional components:

$$x_{et} = [\widetilde{s}_{et}; \widetilde{v}_{et}; \widehat{q}^s_e; \widehat{q}^o_e; W_c \boldsymbol{c}_t + b_c; p_{r_e}] \in \mathbb{R}^{384}, \tag{16}$$

where p_{r_e} comes from a learned 57×64 predicate embedding. An affine/ReLU projection maps to width 256, followed by a 256-channel temporal convolution (kernel five, stride one, zero padding two), ReLU, and a pointwise $256 \to 128 \to 64 \to 1$ MLP. The MLP uses ReLU after its first two layers and a linear scalar output. The convolution operates over the sequence of chronologically ordered sampled times. This produces $z_e(t)$ for Equation 10. Decoder parameters are shared across events and predicates; inference replaces event queries with native-pair queries.

B.2 Controlled baselines

Retained REM. Supervision is restricted to the 3,152 IPS+T or 2,318 VPS events surviving the corresponding tracklet-grounding pipeline's matching, temporal compaction/merging, and final three-frame filtering. IPS+T follows its released visibility rule; VPS requires subject–object co-visibility before merging. The latter exactly reconstructs the saved VPS targets; using the IPS+T rule instead would retain 2,480 events. Full annotated spans are retained for every surviving event. Only count targets and event-wise loss sums change; initialization and fixed class weights are computed from all 3,592 valid events. Architecture, queries, loss coefficients, optimization, and inference remain fixed. The control therefore isolates event deletion rather than truncation. The coverage ablations independently restore deleted events with or without sampled correspondences, as counted in Table 5.

Simple BCE. This baseline uses all valid events and REM's encoders, raw pair logits, temporal context, and interval decoder. Let $\mathcal{G}$ contain eligible pair–time triples (t, i, j) with $(i, j) \in \mathcal{Q}_t$ and valid GT correspondences for both endpoints. Define $\mathcal{N}_r = \mathcal{G} \setminus \bigcup_{e:r_e=r} B_e$. Its local loss is

$$\begin{aligned} \mathcal{L}_{\mathrm{pair}} = -\frac{1}{Z}\Bigg[&\sum_e \frac{\alpha_{r_e}}{\max(1, |B_e|)} \sum_{(t,i,j)\in B_e} \log\sigma(\ell_{r_e tij}) \\ &+ \sum_r \frac{\alpha_r}{\max(1, |\mathcal{N}_r|)} \sum_{(t,i,j)\in\mathcal{N}_r} \log\sigma(-\ell_{rtij}) \Bigg]. \end{aligned} \tag{17}$$

Positive loss is averaged per event and negative loss per predicate; empty sums vanish. Unknown correspondences are excluded from negatives. Predicates are independent, and negative labeling assumes exhaustive relation annotations. Robustness to annotation incompleteness is not evaluated here; this limitation is distinct from matching-induced supervision loss. The full objective is

$$\mathcal{L}_{\mathrm{Simple}} = \mathcal{L}_{\mathrm{count}} + \mathcal{L}_{\mathrm{time}} + \mathcal{L}_{\mathrm{pair}} + \mathcal{L}^{\mathrm{uniform}}_{\mathrm{mark}}, \tag{18}$$

with unit coefficients and the count/time terms from Equation 7. The interval term uses Equation 10 with uniform queries. Events without correspondences retain these global targets but supply no pair-positive or interval loss. The observed/unobserved allocation loss is disabled; its gate remains instantiated but unused, preserving the parameter layout. Primary inference uses uniform queries and pooled sigmoid scores; count/time outputs are auxiliary unless explicitly included in Table 4.

Strong Conv. The same frozen inputs and visibility-masked encoders feed a pair-proposal MLP ($512 \to 1024 \to 1$) with a hidden ReLU. Its inputs concatenate subject and object tracks max-pooled over visible sampled times. Per-time concatenated subject/object features pass through a 512-channel temporal convolution (kernel five, padding two), then ReLU layers of widths 256 and 128. Separate 57-dimensional heads produce temporal and frame-level predicate logits; temporal max pooling gives pair-level predicate logits.

Training retains label-translated assignments; under `legacy_ignore`, annotated positives outside retained spans are ignored by the temporal loss. They are neither restored as positives nor relabeled as negatives. The loss is $\mathcal{L}_{\text{proposal}} + 5\mathcal{L}_{\text{classification}} + \mathcal{L}_{\text{temporal}}$. Proposal/temporal terms use ZLPR, $\log(1 + \sum_{k:y_k=1} e^{-z_k}) + \log(1 + \sum_{k:y_k=0} e^{z_k})$, omitting ignored entries; temporal loss averages only sequences with a sampled positive. Classification uses BCE with positive weight $(\sum_{r'} c_{r'})/c_r$, where c_r is one plus the number of saved assignments for predicate r.

Classification/temporal heads train on at most 50 positive pairs per video, sampled uniformly without replacement when necessary, without additional background pairs. Proposal loss treats unassigned pairs as negatives. Configured crop/pair-frame limits are inactive, and Adam follows the common schedule. Inference proposes at most 100 co-visible pairs by proposal logit d_{ij} and scores

$$s^{\text{Conv}}_{ijr} = \max_{t \in \mathcal{T}} \ell^{\text{Conv}}_{rtij} + \log \sigma(d_{ij}), \tag{19}$$

without an α_r factor. Proposal confidence enters only inference scores and does not modify the predicate-classification training logits. Temporal logits are linearly interpolated and thresholded as for the other local models. This is an architecture reference, not a supervision-matched control.

B.3 Uniform queries and temporal pooling

For a nonempty correspondence or co-observation set B, define the uniform queries

$$q^s_{\text{unif}}(B) = \frac{1}{|B|} \sum_{(t,i,j) \in B} \boldsymbol{s}_{it}, \qquad q^o_{\text{unif}}(B) = \frac{1}{|B|} \sum_{(t,i,j) \in B} \boldsymbol{v}_{jt}. \tag{20}$$

Training uses B_e and inference uses B_{ij}. Every pair–time triple has equal weight, independent of bin width or learned mass. Queries are shared across predicates, but the decoder is predicate-conditioned. Uniform-query REM is trained separately using this aggregation rule at both training and inference, retaining all REM losses and its mass-based score. Simple BCE uses the same uniform queries.

For $p_{rtij} = \sigma(\ell_{rtij})$, let $\mathcal{T}_{ij}$ contain co-visible sampled times, $n_{ij} = |\mathcal{T}_{ij}|$, and $p^{(1)}_{ijr} \geq \cdots \geq p^{(n_{ij})}_{ijr}$ be the sorted probabilities. Pooling uses

$$\begin{aligned} \bar{p}^{\text{mean}}_{ijr} &= \frac{1}{n_{ij}} \sum_{t \in \mathcal{T}_{ij}} p_{rtij}, \quad \bar{p}^{\text{max}}_{ijr} = \max_{t \in \mathcal{T}_{ij}} p_{rtij}, \\ \bar{p}^{\text{top10}}_{ijr} &= \frac{1}{k_{ij}} \sum_{m=1}^{k_{ij}} p^{(m)}_{ijr}, \qquad k_{ij} = \max(1, \lceil 0.1 n_{ij} \rceil). \end{aligned} \tag{21}$$

Invisible times, padding, self-pairs, and pairs without co-observations are excluded. All pooling rules weight sampled observations equally, without temporal-bin weighting. Scores are $\alpha_r \bar{p}^h_{ijr}$, implemented in log space with probability floor 10^{-30}. Each rule reselects up to 100 pairs by maximum predicate score before ranking hypotheses and applying the common interval filter. The pooling comparison reuses one fixed Simple BCE checkpoint per setting and its uniform-query decoder, without retraining. Top-10% is the prespecified primary rule; interval thresholds remain 0 for IPS+T and -0.5 for VPS, without pooling-specific selection.

C Supervision Retention and Spatial Recoverability

C.1 Which events are removed by grounding?

Tracklet-level label translation retains 3,152 of 3,592 valid training events with IPS+T (87.8%) and 2,318 with VPS (64.5%; Table 5). Deleted-event groups have median durations of 5.0–5.4%

Table 5: Training events retained or deleted by tracklet grounding. "Annotated" counts events before excluding out-of-video spans; "Valid" requires nonempty in-video support. Duration is relative to video length. Recoverability uses all frames and a fixed native pair satisfying the strict spatial criterion.

Setting	Event group	Annotated	Valid	Median duration (%)	Recoverable
IPS+T	Retained	3152	3152	42.0	2311
	Deleted, sampled correspondence	109	109	5.0	37
	Deleted, no sampled correspondence	333	331	5.4	20
VPS	Retained	2318	2318	79.0	336
	Deleted, sampled correspondence	488	488	8.6	29
	Deleted, no sampled correspondence	788	786	4.6	7

of the video with IPS+T and 4.6–8.6% with VPS, compared with 42.0% and 79.0% for retained events. Matching therefore disproportionately removes short relations; event normalization prevents duration from determining an event's total temporal-mass supervision. Deletion is distinct from the absence of all usable observations: 109 deleted IPS+T events and 488 deleted VPS events still have sampled correspondences and can supervise the full annotated temporal span. The remaining 331 and 786 valid events contribute count, temporal-mass, and unobserved-state supervision without an interval query. These are the two groups independently restored in Table 3.

C.2 Recoverability and duration

Restoring annotations cannot create missing masks or identities. Table 6 asks whether any fixed native pair can satisfy the strict spatial criterion before ranking and candidate-budget limits. Validation R/mR ceilings are 24.21/30.17 for IPS+T and only 5.03/8.01 for VPS. Among the 1,274 restored VPS events, just 36 are recoverable and 88.1% of normalized mass is unobserved. Sampled correspondence and full-frame recoverability use different grids, so some events without sampled matches remain recoverable over all frames.

Table 6: Spatial recoverability with fixed native tracklet pairs, before relation ranking and candidate-budget limits. The ceilings measure what the supplied tracks can support, not model performance.

	Split	Recoverable / valid	R ceiling (%)	mR ceiling (%)
IPS+T	Train	2368 / 3592	65.92	71.18
	Val	231 / 954	24.21	30.17
VPS	Train	372 / 3592	10.36	14.35
	Val	48 / 954	5.03	8.01

A per-frame GT oracle that can switch tracklet identities raises train/validation recoverability to 2,678/263 events for IPS+T and 2,136/215 for VPS. This diagnostic exposes evidence split across identities; REM does not use oracle merging at inference.

VPS deletes 787/1,100 events shorter than 10% of a video (71.5%), versus 204/1,579 events spanning at least half (12.9%). Figure 3 and Table 7 report duration-group results. For the shortest VPS group, REM improves R/mR over Retained REM from 1.72/2.07 to 2.29/2.61, but both groups shorter than 25% still have zero G-mR. Retaining short-event supervision can improve retrieval without yet achieving accurate interval localization.

D Matched Training and Scoring Controls

D.1 Matched protocol and local losses

REM versus Simple BCE changes several factors jointly. We therefore compare CE and BCE in two training candidate domains, with REM time weights or event-average BCE positive weights. All

Table 7: VPS validation results by annotated event duration at $K = 100$ and vIoU ≥ 0.5. Each model column reports R/mR/G-mR. Predictions are ranked over the full video before GT events are grouped; macro recall averages over predicates represented in each group.

GT duration / video	Events	REM R/mR/G-mR@100	Retained REM R/mR/G-mR@100	Simple BCE R/mR/G-mR@100
$[0, 0.1)$	349	2.29/2.61/0.00	1.72/2.07/0.00	1.72/2.17/0.00
$[0.1, 0.25)$	156	7.05/7.51/0.00	7.05/7.51/0.00	1.92/3.63/0.00
$[0.25, 0.5)$	123	4.88/5.21/1.72	4.88/5.21/0.00	0.81/1.15/0.00
$[0.5, 1]$	326	1.53/2.25/2.25	1.53/2.25/2.25	0.00/0.00/0.00

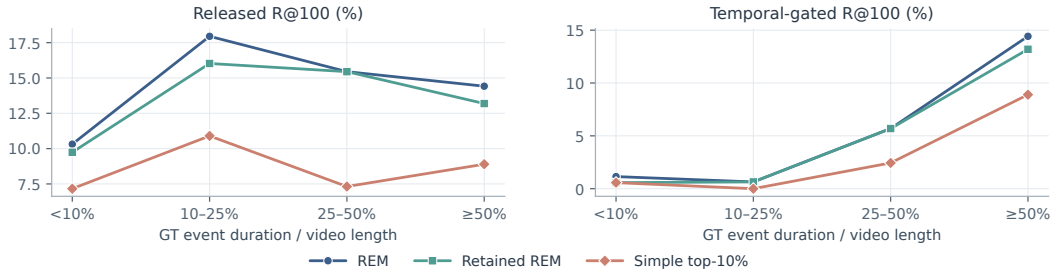

Figure 3: IPS+T relation recall by GT-duration group. Candidate ranking is performed over the full video before GT events are grouped by duration.

six conditions per setting use the same 3,592 events from 338 training videos, frozen observations, architecture, α_r, Z, count, temporal-mass, and observability objectives, complete-span targets, and uniform interval queries. Each has one 100-epoch (1,100-update) training run under the Adam configuration in Appendix B. Only the epoch-100 checkpoint is evaluated, under both readouts below; these are correlated evaluations of 12 fitted models, not 24 independent runs. No checkpoint is selected from these results. Common softmax scoring defines the primary comparisons; sigmoid scoring assesses readout dependence.

At sampled time t, $\mathcal{Q}_t$ contains all ordered pairs of distinct, co-visible predicted tracklets; $\mathcal{G}_t$ is its subset whose two endpoints have unambiguous GT correspondences. The audit verifies $B_{et} \subseteq \mathcal{G}_t \subseteq \mathcal{Q}_t$, with every nonempty B_{et} a singleton. Restricting Q to G preserves all positives and global/interval targets; inference always uses Q. Unknown-correspondence atoms constitute 9.96%/9.45% of the nonpositive Q pool at active predicate–time cells for IPS+T/VPS, using unweighted atom counts. They are not verified semantic negatives; even negatives in G assume annotation completeness.

Define observed positive mass and its predicate–time total by

$$M_{rtij} = \sum_{e:r_e=r} \frac{w_{et} o_{et}}{\max(1, |B_{et}|)} \mathbf{1}[(i,j) \in B_{et}], \qquad O_{rt} = \sum_{i,j} M_{rtij}. \tag{22}$$

For $D \in \{Q, G\}$, let $\mathcal{N}_{rt}^D = \{(i,j) \in \mathcal{D}_t : M_{rtij} = 0\}$ and P^D be the softmax over $\mathcal{D}_t$. The local losses are

$$\mathcal{L}_{\mathrm{CE}}^D = -Z^{-1} \sum_{r,t,i,j} \alpha_r M_{rtij} \log P_{rt}^D(i,j), \tag{23}$$

$$\mathcal{L}_{+}^{\mathrm{time}} = -Z^{-1} \sum_{r,t,i,j} \alpha_r M_{rtij} \log \sigma(\ell_{rtij}), \tag{24}$$

$$\mathcal{L}_{-}^D = -Z^{-1} \sum_{r,t} \alpha_r O_{rt} \frac{\sum_{(i,j)\in\mathcal{N}_{rt}^D} \log \sigma(-\ell_{rtij})}{\max(1, |\mathcal{N}_{rt}^D|)}, \qquad \mathcal{L}_{\mathrm{BCE,time}}^D = \mathcal{L}_{+}^{\mathrm{time}} + \mathcal{L}_{-}^D. \tag{25}$$

Empty negative sets and cells with $O_{rt} = 0$ contribute zero. The negative coefficient is one. With singleton correspondences, Q-CE reproduces uniform-query REM's pair-allocation term, including the 10^{-12} probability floor; this equivalence does not extend to general correspondence bags.

Changing the domain affects both softmax competition and the BCE negative mean, so it tests a candidate-domain policy, not an isolated negative atom.

BCE–avg replaces M only in the positive term by

$$A_{rtij} = \sum_{e:r_e=r} \frac{o_{et}\mathbf{1}[(i,j) \in B_{et}]}{\max(1, \sum_s |B_{es}|)}. \tag{26}$$

The activation mask, O_{rt}, and negative term stay fixed. Each observed event then contributes positive weight one rather than $\rho_e = \sum_t w_{et} o_{et} \leq 1$, increasing class-weighted positive mass by factors 1.488/1.529 for IPS+T/VPS. This intervention changes both positive strength and temporal distribution, not temporal shape alone. Events without correspondences retain global supervision under every condition.

D.2 Common scoring and complete results

All conditions rank with

$$S^{\psi}_{ijr} = \alpha_r C_r \sum_t \pi_{rt} a_{rt} \psi_{rtij}, \qquad \psi_{rtij} = \begin{cases} \mathrm{softmax}_{\mathcal{Q}_t}(\ell_{rt})_{ij}, & \text{softmax readout,} \\ \sigma(\ell_{rtij})\mathbf{1}[(i,j) \in \mathcal{Q}_t], & \text{sigmoid readout.} \end{cases} \tag{27}$$

Ineligible pairs receive zero, as do all pairs when $\mathcal{Q}_t$ is empty. Softmax always uses full Q, including for G-trained models; sigmoid does not define a normalized measure. Each readout independently proposes at most 100 pairs ranked by $\max_r S^{\psi}_{ijr}$, then applies uniform-query interval decoding. Hypotheses whose predicted positive temporal support does not overlap any frame with both endpoints visible are discarded without backfilling. Evaluation follows Appendix E. Learned C, π, a need not coincide across independently trained models. These BCE controls differ from Simple BCE: the observability gate is supervised and all three global factors enter inference.

Table 8: All 12 training conditions under two common readouts: mR@100/G-mR@100 in percent at epoch 100. CE and BCE–time share REM positive weights; BCE–avg changes only the BCE positive term. Q-CE is the uniform-query REM reference. All rows use the global score in Equation 27; BCE bridge rows are not the original Simple BCE baseline.

Training condition	Softmax	Sigmoid
IPS+T		
Q-CE (REM)	14.237/6.624	7.570/3.346
G-CE	14.528/6.598	7.746/3.230
Q-BCE–time	13.284/4.704	14.569/4.561
G-BCE–time	12.018/4.724	13.318/4.951
Q-BCE–avg	13.012/4.538	12.199/4.443
G-BCE–avg	12.946/4.387	12.164/4.362
VPS		
Q-CE (REM)	3.031/0.536	1.962/0.166
G-CE	2.660/0.114	1.962/0.114
Q-BCE–time	2.629/0.485	2.339/0.114
G-BCE–time	2.629/0.485	2.339/0.114
Q-BCE–avg	2.893/0.559	2.079/0.114
G-BCE–avg	2.858/0.523	2.079/0.114

Under common softmax, CE has higher mR point estimates than matched-time BCE in all four settings/domains, but not higher G-mR in VPS/G. For IPS+T/Q, BCE–time exceeds CE under sigmoid scoring (14.569 versus 7.570 mR), reversing their ordering under common softmax. Across readouts, BCE–time with sigmoid also exceeds CE with softmax in mR (14.569 versus 14.237), while CE has higher G-mR (6.624 versus 4.561). This cross-readout comparison is descriptive; the reported CE–BCE confidence intervals apply only to common softmax scoring. With softmax, BCE–avg has higher mR point estimates in three of four comparisons; with sigmoid, BCE–time is higher in all four. These results support configuration-specific comparisons, not universal superiority of CE or REM time weighting.

D.3 Paired uncertainty and interpretation

We use 10,000 paired whole-video bootstrap resamples of the 62 validation videos, pooling hits and GT counts by predicate and averaging recall over predicates with nonzero resampled support (44–54 predicates). Both models in each contrast receive identical video weights. Table 9 reports percentage-point differences with 2.5th/97.5th percentiles using linear interpolation, not averages of per-video mR. Intervals condition on fitted models and evaluation choices; they do not measure training-seed variability or adjust for multiple comparisons or historical validation-set development. Zero endpoints indicate touching zero, not equivalence. For the mR weight–readout interaction, $(\mathrm{time}-\mathrm{avg})_{\mathrm{sigmoid}}-(\mathrm{time}-\mathrm{avg})_{\mathrm{softmax}}$, all four outcomes are resampled jointly; the 95% intervals include or touch zero in all four setting/domain combinations.

Table 9: Matched-control differences under common softmax scoring, in percentage points with paired video 95% confidence intervals. All inference uses full Q. Positive-weight contrasts hold the negative term fixed.

Setting and contrast	ΔmR@100 [95% CI]	ΔG-mR@100 [95% CI]
CE minus matched-time BCE		
IPS+T / Q: CE − BCE–time	$+0.953\ [-2.089, 2.358]$	$+1.920\ [-0.092, 2.334]$
IPS+T / G: CE − BCE–time	$+2.510\ [0.143, 3.515]$	$+1.874\ [-0.108, 2.216]$
VPS / Q: CE − BCE–time	$+0.402\ [-0.254, 1.242]$	$+0.051\ [0.000, 0.170]$
VPS / G: CE − BCE–time	$+0.032\ [-1.253, 0.976]$	$-0.370\ [-1.924, 0.000]$
$Q - G$ training-domain policy		
IPS+T / CE ($Q - G$)	$-0.291\ [-1.064, 0.180]$	$+0.026\ [-0.028, 0.152]$
IPS+T / BCE–time ($Q - G$)	$+1.267\ [-0.110, 2.782]$	$-0.020\ [-0.226, 0.149]$
IPS+T / BCE–avg ($Q - G$)	$+0.066\ [-0.368, 0.630]$	$+0.151\ [0.010, 0.442]$
VPS / CE ($Q - G$)	$+0.370\ [0.000, 1.924]$	$+0.422\ [0.000, 1.975]$
VPS / BCE–time ($Q - G$)	$+0.000\ [0.000, 0.000]$	$+0.000\ [0.000, 0.000]$
VPS / BCE–avg ($Q - G$)	$+0.036\ [0.000, 0.161]$	$+0.036\ [0.000, 0.161]$
BCE time minus event-average positive weights		
IPS+T / Q: time − avg	$+0.272\ [-0.631, 1.692]$	$+0.167\ [-0.489, 1.587]$
IPS+T / G: time − avg	$-0.929\ [-2.646, 0.946]$	$+0.338\ [-0.235, 1.812]$
VPS / Q: time − avg	$-0.265\ [-0.980, 0.000]$	$-0.074\ [-0.260, 0.000]$
VPS / G: time − avg	$-0.229\ [-0.962, 0.130]$	$-0.038\ [-0.229, 0.125]$

The IPS+T/G CE–BCE mR interval excludes zero, whereas most contrasts include or touch zero. VPS differences involve few events: under Q, CE gains three strict hits and loses three across six videos, leaving both models at 27 hits; under G, CE/BCE have 26/27 hits despite CE's slightly higher macro recall. The candidate-domain effect is therefore setting-dependent, and small VPS macro-recall differences should not be read as broad retrieval gains. These controls neither validate unknown pairs as negatives nor isolate improved interval decoding from ranking and candidate filtering.

E Evaluation and Additional Results

E.1 Temporal Localization Diagnostics

Our motivation concerns preserving the *full annotated relation span*, not predicting an arbitrarily long interval. This distinction matters because the released PVSG criterion rewards coverage of the GT span without penalizing predictions outside it. For fixed subject/object masks and categories, its temporal component is

$$Q(p, g) = \frac{\sum_t h_s(t) h_o(t) Y_g(t) \widehat{Y}_p(t)}{D_g}, \tag{28}$$

where h_s and h_o indicate spatial matches, Y_g and $\widehat{Y}_p$ indicate GT and predicted relation activity, and D_g is the GT-valid denominator. Expanding the predicted span cannot decrease this score, even when the additional frames fall outside the annotation.

Table 10: Comparison with SegPVSG with and without RGVA. Best and second-best results among methods are shown in bold and underlined, respectively.

	vIoU ≥ 0.5			vIoU ≥ 0.1		
Method	R/mR@20	R/mR@50	R/mR@100	R/mR@20	R/mR@50	R/mR@100
IPS+T:						
SegPVSG	5.45/9.13	9.54/12.74	13.63/15.91	11.95/16.54	20.13/22.98	28.72/28.92
SegPVSG (w/o RGVA)	5.14/6.67	9.12/10.15	12.58/12.87	12.16/13.20	21.38/21.58	27.57/26.45
REM (Ours)	**8.39**/<u>7.52</u>	**12.05/12.75**	**13.63**/<u>14.28</u>	**18.55**/<u>15.59</u>	**26.42**/<u>22.89</u>	**30.29**/<u>26.78</u>
VPS:						
SegPVSG	1.78/2.26	2.52/2.67	3.04/4.82	9.22/9.01	15.30/15.45	19.29/20.93
SegPVSG (w/o RGVA)	1.78/1.52	2.52/2.16	3.04/2.79	8.39/8.45	15.72/15.33	22.01/21.63
REM (Ours)	**2.10**/<u>1.65</u>	**2.62**/<u>2.24</u>	**3.14**/<u>2.81</u>	**14.47/12.57**	**19.92/17.76**	**24.00/23.35**

Released and temporal metrics. All local results use the 62-video validation split, with 954 relation events over 54 represented predicates. At most 100 native tracklet pairs per video are retained before their 57 predicate hypotheses are ranked. The released strict criterion requires matching subject/object categories and predicate together with the coverage score $Q(p, g) \geq 0.5$ in Equation 28; the weak criterion uses 0.1. To assess localization of the full annotated span rather than coverage alone, G-R and G-mR additionally require

$$\mathrm{tIoU}(p, g) = \frac{|\widehat{S}_p \cap S_g|}{|\widehat{S}_p \cup S_g|} \geq 0.5. \tag{29}$$

Here, $\widehat{S}_p$ and S_g are the predicted and full annotated temporal spans. R@K averages event hits, whereas mR@K first computes recall per predicate and then averages over represented predicates. G-R and G-mR use the same averaging with the additional tIoU condition. GT information is used by the evaluator only after predictions and rankings are fixed.

Replacing every predicted interval with the full video while holding masks, categories, relation scores, and ranking fixed exposes this mismatch: IPS+T mR increases from 14.28 to 17.51, but G-mR falls from 6.75 to 4.71; VPS shows the same direction, from 2.81 to 3.13 mR and from 0.56 to 0.48 G-mR. Higher released recall alone therefore does not demonstrate recovery of the annotated temporal span. The additional tIoU gate and common-candidate tIoU diagnostics assess whether predictions localize the relation rather than merely cover it, which is necessary when evaluating supervision intended to preserve complete temporal annotations.

E.2 Comparison with SegPVSG with and without RGVA.

The main benchmark table reports SegPVSG†, the published variant without relation-centric generative video augmentation (RGVA), to compare methods using the original source annotations. Full SegPVSG additionally synthesizes relation-centric training segments for rare predicates and is reported separately in Table 10. This distinction matters for our motivation: REM seeks to use existing complete annotations despite partial grounding, not to enlarge the training set with generated segments.

The augmented method remains stronger on some metrics. At vIoU ≥ 0.5, full SegPVSG reaches mR@100 of 15.91 with IPS+T and 4.82 with VPS, above REM's 14.28 and 2.81. Conversely, REM obtains higher R@100 at vIoU ≥ 0.1 in both settings (30.29 versus 28.72 and 24.00 versus 19.29). The comparison therefore does not establish uniform superiority over augmented training. It contextualizes gains of REM without additional synthetic data; the retained-event and Simple BCE controls, rather than comparisons with published methods under different training regimes, provide the more direct tests of its supervision formulation.

E.3 Temporal quality and qualitative recovery

Table 11 compares the same spatially eligible candidate–GT tuples (199 IPS+T; 41 VPS), avoiding candidate-set changes as an explanation of decoder quality. REM improves over Retained REM, but

is close to Simple BCE and does not uniformly improve macro tIoU. Thus, retrieval gains cannot be attributed solely to better interval predictions; Section E.1 and Tables 3 and 4 provide complementary fixed-ranking and scoring controls.

Table 11: Temporal-span quality on common spatially eligible candidate–GT tuples.

Setting	Model	Mean tIoU	Macro tIoU
IPS+T	REM	0.373	0.320
	Retained REM	0.361	0.295
	Simple BCE	0.369	0.321
VPS	REM	0.232	0.239
	Retained REM	0.208	0.182
	Simple BCE	0.227	0.216

Figure 4 illustrates two relations recovered only by REM under unchanged native tracks. These examples show relation recovery, not reconstruction of missing masks or identities.

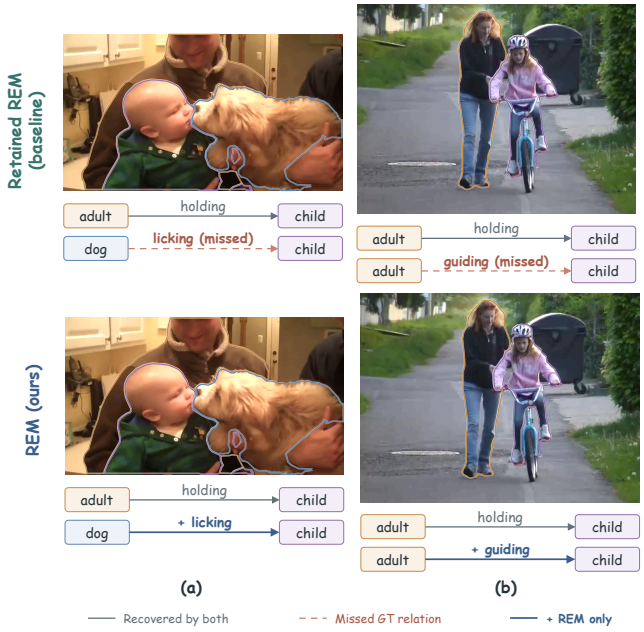

Figure 4: **Qualitative relation recovery.** Top: Retained REM; bottom: REM. Gray denotes relations recovered by both methods, red dashed edges denote GT relations missed by Retained REM, and blue denotes relations additionally recovered by REM. REM recovers *dog–licking–child* in (a) and *adult–guiding–child* in (b).

E.4 Observation-density sensitivity

We reduce the stored observations from at most 128 to 64 uniformly sampled times without retraining or recalibration, keeping the retained masks, class weights, and interval thresholds fixed. Unlike the partition-consistency test in Appendix A.2, this intervention removes visual observations rather than reindexing them, so invariance is not expected. It tests how effectively the frozen models use less temporal evidence.

At 64 samples, REM retains the highest mR and G-mR@100 among the three models in both settings (Table 12), but its IPS+T advantage narrows. REM drops from 14.28 to 11.86 mR and from 6.75 to 4.81 G-mR, whereas Retained REM and Simple BCE change from 10.85 to 10.48 and from 9.54 to 9.09 mR. REM's G-mR margin over Simple BCE consequently shrinks from 1.96 to 0.14 points. With VPS, REM changes from 2.81 to 2.52 mR while retaining 0.56 G-mR, compared with 2.39/0.48

for Retained REM and 1.77/0.05 for Simple BCE. The results show that the advantage survives this particular reduction in sampling, not that REM is insensitive to observation loss. Preserving complete annotations and having sufficient visual evidence remain complementary requirements.

Table 12: Frozen-model sensitivity to temporal sampling at vIoU ≥ 0.5.

	Model	Max. samples	R/mR@20	R/mR@50	R/mR@100	G-mR@100
IPS+T	REM	128	8.39/7.52	12.05/12.75	13.63/14.28	6.75
	REM	64	8.28/7.43	11.84/10.98	13.21/11.86	4.81
	Retained REM	128	8.91/6.86	11.64/9.55	12.68/10.85	4.23
	Retained REM	64	8.70/6.58	11.22/9.10	12.26/10.48	4.03
	Simple BCE	128	1.89/4.83	4.82/7.67	8.39/9.54	4.79
	Simple BCE	64	2.20/5.28	5.45/7.61	8.60/9.09	4.67
VPS	REM	128	2.10/1.65	2.62/2.24	3.14/2.81	0.56
	REM	64	1.89/1.32	2.52/2.21	2.83/2.52	0.56
	Retained REM	128	1.78/1.15	2.62/2.12	2.94/2.44	0.48
	Retained REM	64	1.89/1.22	2.62/2.12	2.83/2.39	0.48
	Simple BCE	128	0.31/0.86	0.52/1.14	1.05/1.90	0.00
	Simple BCE	64	0.21/0.56	0.63/0.99	1.26/1.77	0.05