ADAPTIVE VISUAL EVIDENCE EXTRACTION FOR OPEN-VOCABULARY VIDEO VISUAL RELATIONSHIP DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary video visual relationship detection (Open-VidVRD) aims to recognize objects and their relationships in videos beyond a predefined vocabulary of object and relationship categories. Relationship classification typically relies on representations of trajectory pairs constructed from sampled frames, yet existing pipelines commonly divide videos or candidate relationship intervals into fixed-length temporal segments and sample the same number of frames from each segment. This uniform sampling strategy is suboptimal because relationships exhibit distinct temporal dynamics and require different numbers of frames to capture sufficient visual evidence. Stable relationships may require only a few frames, whereas dynamic ones require more. To address this issue, we propose an adaptive visual evidence extraction architecture for Open-VidVRD. It first partitions videos into temporal segments according to changes in optical flow and bounding box layouts, and then predicts the required number of sampled frames for each segment using aggregated appearance and motion cues. Subsequently, visual evidence is extracted from the selected frames and fused with contextual information from the entire segment for relationship classification. Due to the absence of annotations regarding the optimal number of sampled frames for each segment of a trajectory pair, we construct pseudo supervision by combining relationship-level temporal priors obtained from a multimodal large language model (MLLM) with annotated relationship durations and the temporal coverage of relationship instances within each segment. The MLLM is used only to derive offline supervision and is not invoked during inference. Extensive experiments on the VidVRD and VidOR datasets demonstrate that our method achieves state-of-the-art performance.

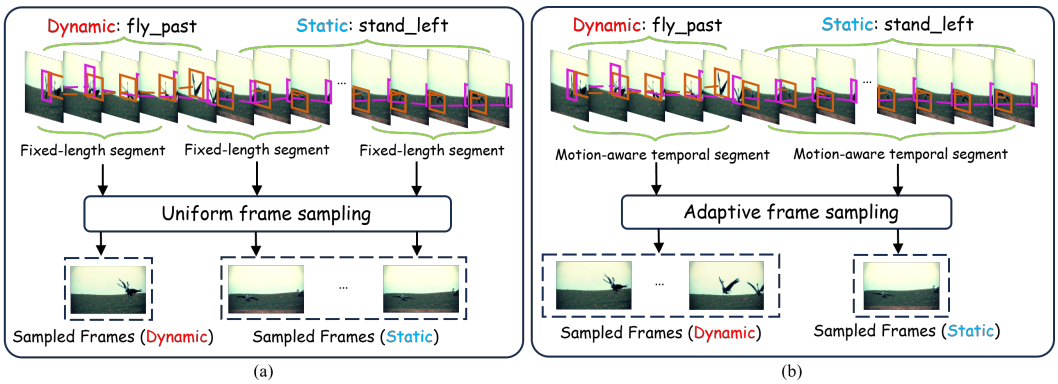

Figure 1: (a) Existing methods use fixed-length temporal segmentation and uniform frame sampling, applying the same observation pattern to different temporal dynamics. (b) Our framework employs motion-aware temporal segmentation and adaptive frame selection to extract visual evidence adaptively.

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

Open-vocabulary video visual relationship detection (Open-VidVRD) Gao et al. (2023) aims to identify relationships between objects in videos without being restricted to a predefined vocabulary of object and relationship categories. In this setting, object and relationship categories are divided into base sets and novel sets. Models are trained using annotations from the base categories and are expected to generalize to unseen novel categories during inference. Since relationships in videos are expressed through interactions between objects, relationship classification typically relies on visual representations built from the visual evidence of trajectory pairs observed across sampled frames.

Existing methods typically divide videos or candidate relationship intervals into fixed-length temporal segments and sample the same number of frames from each segment Gao et al. (2023); Wang et al. (2025a); Yang et al. (2024); Wu et al. (2024). As illustrated in Figure 1(a), this uniform observation pattern overlooks the diverse temporal dynamics of relationships: slowly changing configurations may require only a few observations, while rapidly evolving interactions may require denser visual evidence. Consequently, fixed-frame sampling may preserve redundant views of stable configurations while overlooking critical changes in dynamic intervals. Moreover, fixed segment boundaries may conflate distinct motion patterns or split continuous motion patterns into adjacent segments.

To address this issue, we propose an adaptive visual evidence extraction framework for Open-VidVRD. As illustrated in Figure 1(b), rather than relying on a fixed temporal grid, each video is partitioned into contiguous temporal segments according to changes in global motion, object-region motion, and object layouts. For each trajectory pair, the framework determines the temporal observation requirement for each segment by predicting how many frames to sample from that segment, and then selects the appropriate frames for subsequent visual evidence extraction and relationship classification.

Since the dataset does not annotate how many frames should be sampled for a trajectory pair in each temporal segment, we construct pseudo labels to supervise the temporal observation requirement. Specifically, we first obtain a relationship-level prior from MLLM that indicates the total number of frames required to observe a relationship. For each annotated relationship instance, we then identify the temporal segments covered by its annotated duration and allocate this total number of frames across these segments according to their visual dynamics. The resulting allocation specifies how many frames should be sampled for the relationship in each segment and serves as the pseudo-label for training the observation requirement predictor. This construction is performed offline, and neither relationship annotations nor MLLM-derived priors are used at inference.

To summarize, our main contributions are three-fold:

(1) We propose an adaptive visual evidence extraction framework for Open-VidVRD, which partitions videos into motion-aware temporal segments and adaptively determines the number of frames to be sampled for relationship classification.

(2) We construct pseudo labels specifying the number of frames to sample for each trajectory pair in each temporal segment by combining MLLM-derived relationship-level temporal priors with annotated relationship extents and segment-level visual dynamics, without requiring additional annotations of the temporal observation requirement or MLLM invocation during inference.

(3) We demonstrate the effectiveness of our method on the VidVRD and VidOR benchmarks, achieving state-of-the-art performance while using fewer visual frames for subsequent relationship classification.

2 RELATED WORK

Open-Vocabulary Video Visual Relationship Detection. Open-vocabulary VidVRD (Open-VidVRD) Gao et al. (2023) extends conventional video visual relationship detection (VidVRD) Shang et al. (2017) to object and predicate categories unseen during training. VidVRD localizes subject–predicate–object relationships that evolve over time, supporting various video understanding applications Zhao et al. (2023); Nguyen et al. (2024); Rodin et al. (2024). Existing methods primarily focus on improving cross-modal alignment of trajectory-pair representations and general-

ization capabilities for open categories. RePro Gao et al. (2023) learns compositional relationship prompts, while OV-MMP Yang et al. (2024) and UASAN Wu et al. (2024) strengthen visual–textual correspondence through multimodal prompting and semantic alignment. EOV-MMP Wang et al. (2025a) further integrates open-vocabulary representation learning with relationship classification into an end-to-end framework. While these methods improve the alignment between visual relation evidence and linguistic semantics, they typically operate on frames provided by predefined sampling processes. In contrast, our method adaptively determines how many frames to sample and which representative frames to retain for each trajectory pair prior to detailed trajectory-pair encoding.

Adaptive Temporal Evidence Extraction. Adaptive temporal evidence extraction aims to acquire informative visual observations without relying on a fixed temporal observation scheme. Existing methods select observations based on relevance, diversity, or intermediate semantic information. Correlation- and relevance-based approaches exploit object grounding and confidence Ye et al. (2025), visual clustering and language alignment Park et al. (2026), or temporal correlation and visual diversity with hierarchical organization Liu et al. (2025); Wang et al. (2025b). Inference-based approaches further leverage intermediate predictions, confidence, search, or verification to identify relevant frames Fan et al. (2025a;b); Guo et al. (2025), primarily for query-conditioned video reasoning. In contrast, our framework is query-free and predicate-agnostic at inference. For each trajectory pair-segment combination, it predicts the number of frames to be sampled based on appearance and motion dynamics, and then selects representative frames based on quality and visual diversity. Thus, rather than only determining which frames to retain, our framework jointly determines how many frames are needed and which frames to select for open-vocabulary relationship classification.

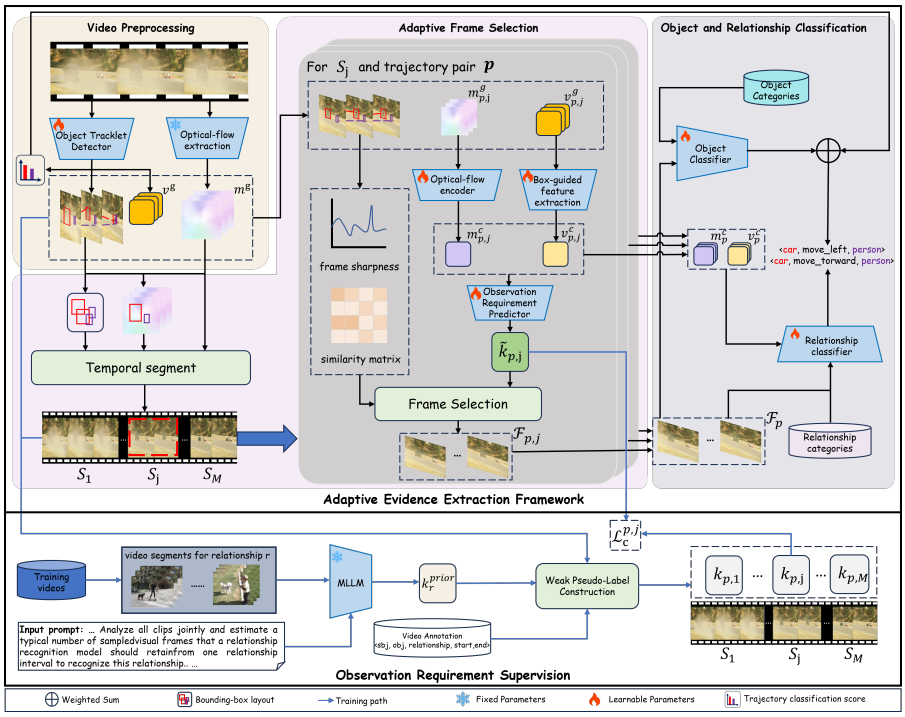

Figure 2: Overview of our proposed method.

3 METHOD

3.1 OVERVIEW

We propose an adaptive visual evidence extraction framework consisting of three main modules: video preprocessing, adaptive frame selection, and object and relationship classification, where adaptive frame selection is our main contribution. The video preprocessing module adopts an object

tracklet detector and an optical-flow extractor to obtain object trajectories, visual features, and motion representations. Based on these representations, the adaptive frame selection module partitions the video into motion-aware temporal segments, predicts the observation requirement for each trajectory pair within each segment, and selects representative frames accordingly. Finally, the object and relationship classification module aggregates the selected frames and their segment-level motion and appearance features to produce the final object and relationship predictions.

Since existing datasets do not annotate how many frames should be sampled for each trajectory pair within a temporal segment, we introduce observation requirement supervision to construct weak training targets from MLLM-derived relationship-level temporal priors, annotated relationship extents, and segment-level visual dynamics. These targets supervise the observation requirement predictor during training, while neither the MLLM-derived priors nor the relationship annotations are required at inference. An overview of our method is illustrated in Figure 2.

3.2 Adaptive Visual Evidence Extraction Framework

Video Preprocessing. Given an input video V, we first adopt an object tracklet detector and an optical-flow extraction module to obtain the basic representations required by our framework. The object tracklet detector produces a set of object trajectories $\mathcal{T} = \{T^n\}_{n=1}^{N_T}$ and their corresponding trajectory-level object classification scores $\{\mathbf{s}_n^{\text{traj}}\}_{n=1}^{N_T}$. It also provides a global visual feature $\mathbf{v}_{f_t}^g$ for each frame f_t, forming $\mathbf{v}^g = \{\mathbf{v}_{f_t}^g\}_{t=1}^{L}$. In parallel, the optical-flow extraction module computes $\mathbf{m}_{f_t}^g$ from the optical flow between f_{t-1} and f_t, yielding $\mathbf{m}^g = \{\mathbf{m}_{f_t}^g\}_{t=2}^{L}$. These modules are used only to provide trajectory, appearance, and motion representations for the subsequent adaptive evidence extraction process.

Adaptive Frame Selection. We first use the precomputed motion representations $\mathbf{m}^g$ and object trajectories $\mathcal{T}$ to partition the video into motion-aware temporal segments. The normalized temporal changes in global motion, object-region motion, and object layouts at timestamp t are denoted by $\widehat{\Delta}_g(\mathbf{m}^g; t)$, $\widehat{\Delta}_m(\mathbf{m}^g, \mathcal{T}; t)$, and $\widehat{\Delta}_l(\mathcal{T}; t)$, respectively. We compute the boundary score as

$$D_t = \alpha_g \widehat{\Delta}_g(\mathbf{m}^g; t) + \alpha_m \widehat{\Delta}_m(\mathbf{m}^g, \mathcal{T}; t) + \alpha_l \widehat{\Delta}_l(\mathcal{T}; t). \tag{1}$$

Local maxima of D_t are treated as candidate boundaries and refined using temporal non-maximum suppression, yielding M contiguous segments $\mathcal{S} = \{S_j\}_{j=1}^{M}$. These segments are shared by all candidate trajectory pairs and serve as temporal units for adaptive evidence allocation, rather than semantic events or relationship boundaries.

For each candidate subject–object trajectory pair $p = (T^s, T^o)$ and segment S_j, we construct the candidate frame set

$$\mathcal{C}_{p,j} = \{t \in S_j \mid b_t^s \in T^s,\ b_t^o \in T^o\}. \tag{2}$$

Pair–segment instances with empty candidate sets are discarded. For each valid instance, an optical-flow encoder and a box-guided feature extractor operate on the corresponding segment and trajectory pair to produce pair-specific motion and appearance features:

$$\mathbf{m}_{p,j}^c = E_{\text{flow}}\left(\mathbf{m}_{S_j}^g, T^s, T^o\right), \qquad \mathbf{v}_{p,j}^c = E_{\text{box}}\left(\mathbf{v}_{S_j}^g, T^s, T^o\right), \tag{3}$$

where $\mathbf{m}_{S_j}^g$ and $\mathbf{v}_{S_j}^g$ denote the optical-flow and global visual representations within S_j, respectively. These pair-specific features are used to predict the continuous observation requirement:

$$\widetilde{k}_{p,j} = g_{\text{reg}}\left([\mathbf{v}_{p,j}^c; \mathbf{m}_{p,j}^c]\right). \tag{4}$$

We then obtain a feasible integer observation count by rounding the prediction and constraining it to the valid sampling range:

$$\widehat{k}_{p,j} = \text{clip}\left(\text{round}(\widetilde{k}_{p,j}), 1, \min(K_{\max}, |\mathcal{C}_{p,j}|)\right). \tag{5}$$

We then select $\widehat{k}_{p,j}$ representative frames from $\mathcal{C}_{p,j}$. Given a candidate frame t and the already selected set $\mathcal{A}$, we define

$$Q_{p,j}(t \mid \mathcal{A}) = \eta_c c_{p,j}(t) + \eta_s s_{p,j}(t) - \lambda_{\text{red}} \max_{t' \in \mathcal{A}} \text{sim}\left(\mathbf{v}_{f_t}^g, \mathbf{v}_{f_{t'}}^g\right), \tag{6}$$

where $c_{p,j}(t)$ measures the spatial separation between the subject and object trajectory boxes, $s_{p,j}(t)$ measures frame sharpness, and $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity between frame-level global visual features. We greedily select $\widehat{k}_{p,j}$ frames according to $Q_{p,j}$, favoring frames with high spatial coverage and sharpness while reducing visual redundancy. The resulting frame set is denoted by $\mathcal{F}_{p,j}$.

Object and Relationship Classification. For a trajectory pair p, we concatenate the representative frames selected from all temporal segments in chronological order: $\mathcal{F}_p = \mathrm{Concat}_{j=1}^{M}\left(\mathcal{F}_{p,j}\right)$. The corresponding segment-level motion and appearance features are aggregated in the same temporal order to form pair-level representations:

$$\mathbf{m}_p^c = \Phi_m\left(\{\mathbf{m}_{p,j}^c\}_{j=1}^{M}\right), \qquad \mathbf{v}_p^c = \Phi_v\left(\{\mathbf{v}_{p,j}^c\}_{j=1}^{M}\right), \tag{7}$$

where Φ_m and Φ_v temporally aggregate the motion and appearance features, respectively, while segments without candidate frames are omitted. The selected frames in $\mathcal{F}_p$, together with $\mathbf{m}_p^c$ and $\mathbf{v}_p^c$, are encoded by the downstream visual, geometric, motion, and optical-flow modules, followed by spatial and temporal decoders.

We further perform object classification on the selected frames and aggregate the resulting scores across the temporal segments to obtain $\mathbf{s}_p^{\mathrm{seg}}$. Let $\mathbf{s}_p^{\mathrm{traj}}$ denote the trajectory-level object classification scores for the subject and object trajectories in p. We fuse the two sources of object evidence as

$$\mathbf{s}_p^{\mathrm{obj}} = \lambda_{\mathrm{obj}}\mathbf{s}_p^{\mathrm{traj}} + (1-\lambda_{\mathrm{obj}})\mathbf{s}_p^{\mathrm{seg}}, \tag{8}$$

where λ_{obj} controls the relative contributions of trajectory-level and selected-frame predictions. Finally, the selected visual evidence and the pair-level contextual representations $\mathbf{m}_p^c$ and $\mathbf{v}_p^c$ are fused and compared with open-vocabulary predicate embeddings to obtain multi-label predicate scores. Further implementation details of our method are provided in Appendix A.

3.3 Observation Requirement Supervision

For each base relationship category r used during training, we select a limited number of training clips containing annotated instances of r and denote them by $\mathcal{V}_r$. We provide these clips, together with the category r and a fixed prompt $\mathcal{P}$, to an MLLM and ask it to jointly estimate how many frames are generally required to observe the temporal evidence of r. The resulting relationship-level temporal prior is

$$k_r^{\mathrm{prior}} = \mathrm{MLLM}\left(\mathcal{V}_r, r; \mathcal{P}\right). \tag{9}$$

This prior provides a coarse relationship-level estimate of the required observation budget and is used only for offline pseudo-label construction.

Let $\mathcal{R}$ denote the set of annotated relationship instances. For each instance $\rho \in \mathcal{R}$, let r_ρ, $p(\rho)$, and τ_ρ denote its relationship category, subject–object trajectory pair, and temporal extent, respectively. We retain the temporal segments that overlap with τ_ρ and contain available candidate frames:

$$\mathcal{S}_\rho^+ = \left\{S_j \in \mathcal{S} \mid S_j \cap \tau_\rho \neq \emptyset,\ |\mathcal{C}_{p(\rho),j}| > 0\right\}. \tag{10}$$

Using the optical-flow and trajectory-layout representations obtained during preprocessing, we compute the visual dynamics and allocation weight of each valid segment as

$$d_{\rho,j} = \alpha_{\mathrm{alloc}}\widehat{d}_{\rho,j}^{\,\mathrm{flow}} + (1-\alpha_{\mathrm{alloc}})\widehat{d}_{\rho,j}^{\,\mathrm{layout}}, \qquad w_{\rho,j} = \frac{\epsilon + d_{\rho,j}}{\sum_{S_\ell \in \mathcal{S}_\rho^+}(\epsilon + d_{\rho,\ell})}. \tag{11}$$

We determine the total observation budget for ρ by combining its relationship-level prior with the available candidate-frame capacities:

$$K_{\rho,j}^{\mathrm{cap}} = \min\left(K_{\max}, |\mathcal{C}_{p(\rho),j}|\right), \qquad \bar{k}_\rho = \min\left(\sum_{S_j \in \mathcal{S}_\rho^+} K_{\rho,j}^{\mathrm{cap}}, \max\left(|\mathcal{S}_\rho^+|, k_{r_\rho}^{\mathrm{prior}}\right)\right). \tag{12}$$

The budget $\bar{k}_\rho$ is allocated across the valid segments according to $\{w_{\rho,j}\}$ using a capacity-constrained largest-remainder procedure, producing segment-level targets $k_{\rho,j}$. This procedure assigns at least one observation to each valid segment and allocates more observations to segments with greater visual dynamics while respecting their capacities.

If multiple annotated relationship instances correspond to the same trajectory pair and segment, their allocated targets are not accumulated. We instead define

$$\mathcal{R}_{p,j} = \{\rho \in \mathcal{R} \mid p(\rho) = p,\ S_j \cap \tau_\rho \neq \emptyset\} \tag{13}$$

and use the largest allocated target:

$$K_{p,j}^{\text{cap}} = \min\left(K_{\max}, |\mathcal{C}_{p,j}|\right), \qquad k_{p,j} = \min\left(K_{p,j}^{\text{cap}}, \max_{\rho \in \mathcal{R}_{p,j}} k_{\rho,j}\right). \tag{14}$$

The valid training set is defined as

$$\mathcal{M} = \{(p,j) \mid \mathcal{R}_{p,j} \neq \emptyset \ \wedge\ |\mathcal{C}_{p,j}| > 0\}. \tag{15}$$

For each $(p,j) \in \mathcal{M}$, the resulting $k_{p,j}$ serves as the weak target for the predicted observation requirement $\widetilde{k}_{p,j}$. Relationship annotations and MLLM-derived priors are used only during training; neither the MLLM nor the annotation-dependent pseudo-label construction is required at inference. Additional details of the MLLM prompt, input construction, and capacity-constrained target allocation are provided in Appendix B.

3.4 Training Objective

We adopt a two-stage training strategy. In the first stage, the observation requirement predictor, optical-flow encoder and box-guided feature extraction are trained on the valid pair–segment samples in $\mathcal{M}$ using the pseudo labels:

$$\mathcal{L}_{\text{c}} = \frac{1}{|\mathcal{M}|} \sum_{(p,j)\in\mathcal{M}} \mathcal{L}_c^{p,j} = \frac{1}{|\mathcal{M}|} \sum_{(p,j)\in\mathcal{M}} \text{SmoothL1}\left(\widetilde{k}_{p,j}, k_{p,j}\right). \tag{16}$$

In the second stage, the observation requirement predictor, optical-flow encoder, box-guided feature extraction and the relationship classifier are jointly optimized. The greedy frame-selection procedure described above induces an ordered sequence of candidate frames for each trajectory pair within a segment, where i denotes the rank in this sequence. To approximate retaining the first $\widehat{k}_{p,j}$ frames while preserving gradients with respect to the continuous prediction $\widetilde{k}_{p,j}$, we employ a straight-through cumulative mask:

$$m_{p,j,i}^{\text{soft}} = \sigma\left(\frac{\widetilde{k}_{p,j} - i + \delta}{\gamma}\right), \qquad m_{p,j,i}^{\text{ST}} = \mathbf{1}\left[i \leq \widehat{k}_{p,j}\right] - \text{sg}\left(m_{p,j,i}^{\text{soft}}\right) + m_{p,j,i}^{\text{soft}}, \tag{17}$$

where γ controls the softness of the approximation, δ aligns the soft and discrete thresholds, and $\text{sg}(\cdot)$ denotes the stop-gradient operation. The hard mask determines the frames retained in the forward pass, while the soft mask provides an approximate gradient for $\widetilde{k}_{p,j}$. We treat the greedy frame order as fixed during backpropagation. The masked frame representations are then used for relationship classification, allowing the relationship loss to refine the predicted observation requirement based on the usefulness of the selected evidence. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{rel}} + \lambda_{\text{c}} \mathcal{L}_{\text{c}}, \tag{18}$$

where $\mathcal{L}_{\text{rel}}$ denotes the multi-label predicate classification loss and λ_{c} balances the pseudo-label supervision. Thus, $\mathcal{L}_{\text{c}}$ provides an initial learning signal for observation-count prediction, while $\mathcal{L}_{\text{rel}}$ further refines the predictor according to the visual evidence required for relationship recognition.

4 Experiment

Datasets and Evaluation Protocol. We evaluate our method on the VidVRD Shang et al. (2017) and VidOR Shang et al. (2019) datasets. VidVRD contains 1,000 videos, with 800 for training and 200 for testing, covering 35 object categories and 132 predicate categories. VidOR contains 10,000 videos, with 7,000 videos for training, 835 for validation, and 2,165 for testing, covering 80 object categories and 50 predicate categories. Following RePro Gao et al. (2023), we use a frequency-based base/novel split, treating common and rare categories as base and novel, respectively. We train on the base categories and evaluate on the Novel and All splits using SGDet, SGCls, and PredCls on the VidVRD test set and VidOR validation set, since VidOR test annotations are unavailable. We report relationship mAP and R@50/100, together with object trajectory mAP_o following EOV-MMP Wang et al. (2025a).

Table 1: Open-vocabulary video visual relationship detection results on VidVRD.

Split	Method	SGDet				SGCls				PredCls		
		mAP	R@50	R@100	mAP_o	mAP	R@50	R@100	mAP_o	mAP	R@50	R@100
Novel	ALPro	0.98	2.79	4.33	10.36	3.69	7.27	8.92	21.06	4.09	9.42	10.41
	VidVRD-II	3.11	7.93	11.38	10.36	5.70	13.22	18.34	21.06	7.35	18.84	26.44
	RePro	5.87	12.75	16.23	10.36	10.32	19.17	25.28	21.06	12.74	25.12	33.88
	OV-MMP	12.15	13.72	15.21	14.37	17.57	21.98	28.43	24.96	21.14	30.41	37.85
	EOV-MMP	15.04	16.03	18.18	36.31	17.96	30.74	36.86	31.73	21.65	35.37	43.64
	Ours	**18.08**	**17.12**	**19.76**	**40.41**	**20.81**	**32.61**	**38.88**	**42.32**	**23.09**	**36.06**	**45.17**
All	ALPro	3.03	2.57	3.11	18.18	3.92	3.88	4.75	68.99	4.97	4.50	5.79
	VidVRD-II	12.66	9.72	12.50	18.18	17.26	14.93	19.68	68.99	19.73	18.17	24.90
	RePro	21.12	12.63	15.42	18.18	30.15	19.75	25.00	68.99	34.90	25.50	32.49
	OV-MMP	22.10	13.26	16.08	34.61	29.38	23.56	28.89	70.39	38.08	30.47	37.46
	EOV-MMP	26.34	16.48	19.54	52.72	31.95	25.96	31.66	74.25	39.83	31.66	39.69
	Ours	**28.93**	**17.33**	**20.86**	**56.23**	**34.82**	**27.16**	**33.47**	**81.48**	**41.22**	**32.99**	**41.46**

Table 2: Open-vocabulary video visual relationship detection results on VidOR.

Split	Method	SGDet				SGCls				PredCls		
		mAP	R@50	R@100	mAP_o	mAP	R@50	R@100	mAP_o	mAP	R@50	R@100
Novel	ALPro	-	-	-	-	-	3.17	3.74	-	-	8.35	9.79
	VidVRD-II	-	-	-	-	-	1.44	2.01	-	-	4.32	4.89
	RePro	-	-	-	-	-	2.01	2.30	-	-	7.20	8.35
	OV-MMP	0.84	1.44	1.44	1.11	2.40	5.48	6.92	6.04	3.58	9.22	11.53
	EOV-MMP	2.45	4.79	4.79	2.33	2.72	5.76	8.65	6.83	4.11	9.80	14.41
	Ours	**4.66**	**5.96**	**5.96**	**5.09**	**5.25**	**6.17**	**9.98**	**9.72**	**5.08**	**10.22**	**15.62**
All	ALPro	-	-	-	-	-	0.95	1.32	-	-	2.61	3.66
	VidVRD-II	-	-	-	-	-	9.40	12.78	-	-	24.81	34.11
	RePro	-	-	-	-	-	10.03	12.91	-	-	27.11	35.76
	OV-MMP	7.15	6.54	8.29	3.38	24.00	23.04	30.14	25.86	38.52	33.44	43.80
	EOV-MMP	11.08	8.43	9.82	12.99	25.21	23.78	30.17	26.17	39.75	33.68	43.87
	Ours	**13.17**	**9.45**	**10.54**	**15.50**	**26.65**	**24.96**	**30.33**	**28.94**	**41.47**	**34.19**	**45.03**

Implementation Details. We adopt the ViT-L/14@336px variant of CLIP as the visual encoder and keep its parameters fixed. The maximum number of frames sampled from each trajectory-pair–segment combination is set to $K_{\max} = 24$. For motion-aware temporal segmentation, the boundary-score weights are set to $\alpha_g = \alpha_m = \alpha_l = 1/3$. The flow-layout allocation weight is set to $\alpha_{\text{alloc}} = 0.6$. For representative frame selection, the weights of the spatial-separation and sharpness terms are set to $\eta_c = 0.4$ and $\eta_s = 0.6$, respectively, with the redundancy penalty set to $\lambda_{\text{red}} = 0.5$. For trajectory classification, the trajectory-level and segment-level object classification scores are fused with $\lambda_{\text{obj}} = 0.6$. All experiments are conducted on a single NVIDIA RTX 4090D GPU with 24 GB memory. Additional experimental settings are provided in Appendix C.

4.1 Comparison Results

We compare our method with existing Open-VidVRD methods, including ALPro Li et al. (2022), VidVRD-II Shang et al. (2021), RePro Gao et al. (2023), OV-MMP Yang et al. (2024), and EOV-MMP Wang et al. (2025a). For all compared methods, the baseline results reported in Tables 1 and 2 are directly taken from the evaluation reported by EOV-MMP, rather than from the original papers or independent reruns in this work. Following EOV-MMP, these results were obtained under a consistent open-vocabulary evaluation protocol using trajectory detectors retrained after excluding novel-category data. Results for ALPro, VidVRD-II, and RePro on VidOR SGDet are unavailable in the EOV-MMP evaluation and are therefore omitted.

As shown in Tables 1 and 2, our method achieves the best SGDet performance on both the Novel and All splits of VidVRD and VidOR. On VidVRD, our method achieves 18.08% and 28.93% mAP on the Novel and All splits, respectively, surpassing EOV-MMP by 3.04% and 2.59%, respectively. It also improves R@50, R@100, and mAP_o by up to 1.09%, 1.58%, and 4.10%, respectively. On VidOR, our method achieves 4.66% and 13.17% mAP on the Novel and All splits, respectively, exceeding EOV-MMP by 2.21% and 2.09%, respectively. The corresponding improvements in R@50, R@100, and mAP_o reach up to 1.17%, 1.17%, and 2.76%, respectively. These consistent gains on

Table 3: Component analysis on VidVRD. Seg.: segmentation; Div.: diversity.

Variant	Novel		All	
	mAP_o	mAP	mAP_o	mAP
Baseline	34.14	14.45	49.32	25.22
+ Count Prediction	34.14	16.65	49.32	26.36
+ Motion-aware temporal Seg.	34.14	17.15	49.32	27.31
+ Quality-Div. Selection	34.14	17.73	49.32	27.92
+ Segment Context	34.14	17.99	49.32	28.35
Full Model	**40.41**	**18.08**	**56.23**	**28.93**

Table 4: Ablation of observation requirement supervision on VidVRD. Sup.: supervision.

Method	Avg. Frames↓	Novel mAP	All mAP
No Count Sup.	14.54	16.15	25.63
Uniform Prior	3.82	16.44	26.56
Duration Prior	3.79	16.95	27.10
MLLM Prior	3.68	17.34	27.62
Pseudo-Target Sup.	3.56	17.95	27.85
Full Model	**3.12**	**18.08**	**28.93**

Table 5: Adaptive observation-count allocation versus fixed-budget sampling on VidVRD.

Method	Avg. Frames↓	Novel mAP	All mAP
Uniform Sampling	4	17.42	27.39
Random Sampling	8	15.34	25.87
Attention-based Sampling	4	17.32	26.56
A.I.R.-based Sampling	4	14.49	27.04
ReaSon-based Sampling	8	15.38	27.26
AKS-based Sampling	4	14.12	26.46
Ours	**3.12**	**18.08**	**28.93**

Table 6: Sampling behavior across relationships with different temporal dynamics on VidVRD.

Method	Dynamic		Static	
	Avg. Frames↓	mAP	Avg. Frames↓	mAP
Baseline	**5.19**	28.83	5.63	29.77
Ours	6.42	**32.44**	**2.49**	**29.98**

both Novel and All splits demonstrate the effectiveness of adaptive visual evidence extraction for open-vocabulary relationship detection.

4.2 Ablation Studies

All ablation studies are conducted on the VidVRD test set under the SGDet setting. We use an EOV-MMP-based baseline, which divides each video into fixed-length 30-frame segments and samples one frame from each segment. In addition to relationship mAP, we report Avg. Frames, defined as the average number of selected frames per trajectory pair.

Component Analysis. We progressively add the main components cumulatively, as reported in Table 3. Count prediction is first introduced under fixed-length segmentation, followed by motion-aware segmentation, quality-diversity frame selection, and segment context aggregation. The Full Model further adds trajectory/selected-frame object-score fusion, which mainly improves object detection, confirming the complementary roles of adaptive evidence extraction and object refinement.

Observation-Count Supervision. Table 4 evaluates different supervision strategies for learning the observation requirement. Relationship-level temporal priors improve both detection performance and frame efficiency, with MLLM-derived priors outperforming simpler duration-based priors. Adding segment-level dynamics through pseudo-target supervision further improves the results, while joint optimization with relationship classification achieves the best performance with the fewest selected frames. Further ablation results on temporal segmentation are reported in Appendix D.1.

Adaptive Observation-Count Allocation versus Fixed Budgets. We compare our adaptive observation-count allocation strategy with uniform and random sampling, as well as existing frame-selection methods including A.I.R. Zou et al. (2026), ReaSon Zhou et al. (2026), and AKS Tang et al. (2025), as reported in Table 5. The compared methods use a predefined or fixed sampling budget, whereas our method predicts the observation count for each trajectory pair within each temporal segment. Even when the competing methods are evaluated with their best-performing budgets, our method achieves higher Novel and All mAP while selecting fewer frames, demonstrating the benefit of dynamics-aware observation-count allocation. Further ablation results on the maximum sampling budget are reported in Appendix D.2.

Table 7: Efficiency comparison on the VidVRD test set.

Method	Avg. Frames ↓	FLOPs / Video ↓	Runtime / Video ↓	Novel mAP	All mAP
Baseline	5.45	14.532 TFLOPs	612 ms	16.45	25.22
Ours	**3.12**	**11.762 TFLOPs**	**509 ms**	**18.08**	**28.93**

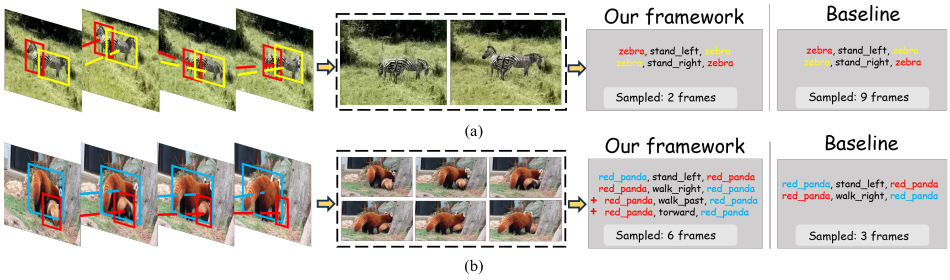

Figure 3: Qualitative comparison of frame selection between our framework and the baseline under different temporal dynamics. The "+" symbol indicates additional relationship triplets correctly predicted by our framework but missed by the baseline.

Adaptive Allocation across Visual Dynamics. Table 6 analyzes the sampling behavior of our observation-count predictor for samples with different levels of visual dynamics. Our method allocates more frames to samples exhibiting stronger temporal changes and fewer frames to relatively static samples, rather than uniformly reducing the sampling budget. This result demonstrates that the predicted observation count responds to the observed motion and layout variations.

4.3 Efficiency Analysis

We compare our method with the baseline under identical hardware, evaluation settings, trajectory candidates, and pre-extracted frame-level features. The costs of trajectory detection and feature extraction are excluded because they are shared by all methods. As shown in Table 7, our method selects fewer frames and reduces FLOPs and inference time while achieving better relationship detection performance, demonstrating the efficiency of adaptive visual evidence extraction.

4.4 Qualitative Analysis

Figure 3 illustrates how our method adaptively selects frames according to the temporal dynamics of different relationships. In (a), the relationship exhibits relatively consistent visual evidence over time, allowing our method to retain only a few representative frames while using substantially fewer frames than the baseline. In contrast, (b) involves more diverse visual changes, for which our method selects additional frames to capture sufficient temporal evidence and consequently identifies more relationship triplets than the baseline. These examples demonstrate that our method does not uniformly increase or reduce the number of sampled frames, but dynamically adjusts frame selection according to the temporal characteristics of each relationship. Additional qualitative comparisons and failure cases are provided in Appendix E.

5 Conclusion

In this work, we propose an adaptive visual evidence extraction framework for Open-VidVRD that dynamically segments videos and determines the number of frames to sample for each trajectory-pair segment. We construct pseudo labels by combining MLLM-derived relationship-level temporal priors with relationship extents and segment-level dynamics, using the MLLM only during training. Experiments on VidVRD and VidOR demonstrate state-of-the-art performance while reducing redundant frames for relationship classification.

AI USE STATEMENT

In this work, we used generative AI tools for limited assistance in improving the readability, grammar, and presentation of the manuscript, as well as for assisting with software code debugging. We also use MLLM as a component of the proposed method to obtain relationship-level temporal priors for constructing offline pseudo supervision. The MLLM-generated priors are only used during training for pseudo-label construction and are not involved in inference.

We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for mathematical proofs, write proofs, or replace the authors' decisions in designing the research methodology, experiments, or analysis. The proposed framework, experimental protocols, implementation details, and conclusions were designed and verified by the authors.

All AI-assisted code was reviewed, modified, and tested by the authors to ensure correctness. All AI-assisted text was carefully reviewed and revised by the authors for accuracy, consistency, and fidelity to the actual work. The authors take responsibility for the final content of this work, including all text, claims, experimental results, and artifacts produced with the aid of generative AI.

REFERENCES

Sunqi Fan, Meng-Hao Guo, and Shuojin Yang. Agentic keyframe search for video question answering, 2025a. URL https://arxiv.org/abs/2503.16032.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 75–92, Cham, 2025b. Springer Nature Switzerland. ISBN 978-3-031-72670-5.

Kaifeng Gao, Long Chen, Hanwang Zhang, Jun Xiao, and Qianru Sun. Compositional prompt tuning with motion cues for open-vocabulary video relation detection, 2023. URL https://arxiv.org/abs/2302.00268.

Weiyu Guo, Ziyang Chen, Shaoguang Wang, JianXiang He, Yijie Xu, Jinhui Ye, Ying Sun, and Hui Xiong. Logic-in-frames: Dynamic keyframe search via visual semantic-logical verification for long video understanding. *ArXiv*, abs/2503.13139, 2025. URL https://api.semanticscholar.org/CorpusID:277103913.

Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven C.H. Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4943–4953, 2022. doi: 10.1109/CVPR52688.2022.00490.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. BOLT: Boost Large Vision-Language Model Without Training for Long-Form Video Understanding . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3318–3327, Los Alamitos, CA, USA, June 2025. IEEE Computer Society. doi: 10.1109/CVPR52734.2025.00315. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.00315.

Trong-Thuan Nguyen, Pha Nguyen, and Khoa Luu. Hig: Hierarchical interlacement graph approach to scene graph generation in video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18384–18394, 2024. doi: 10.1109/CVPR52733.2024.01740.

Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryu, Donghyun Kim, and Michael S. Ryoo. Too many frames, not all useful: Efficient strategies for long-form video qa, 2026. URL https://arxiv.org/abs/2406.09396.

Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, and Giovanni Maria Farinella. Action scene graphs for long-form understanding of egocentric videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18622–18632, 2024. doi: 10.1109/CVPR52733.2024.01762.

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*, MM '17, pp. 1300–1308, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349062. doi: 10.1145/3123266.3123380. URL https://doi.org/10.1145/3123266.3123380.

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. ACM, 2019.

Xindi Shang, Yicong Li, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Video visual relation detection via iterative inference. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pp. 3654–3663, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386517. doi: 10.1145/3474085.3475263. URL https://doi.org/10.1145/3474085.3475263.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128. IEEE, 2025.

Yongqi Wang, Xinxiao Wu, Shuo Yang, and Jiebo Luo. End-to-end open-vocabulary video visual relationship detection using multi-modal prompting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6599–6615, 2025a. doi: 10.1109/TPAMI.2025.3561366.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive Tree-based Video Representation for LLM Reasoning on Long Videos . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3272–3282, Los Alamitos, CA, USA, June 2025b. IEEE Computer Society. doi: 10.1109/CVPR52734.2025.00311. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.00311.

Ziyue Wu, Junyu Gao, and Changsheng Xu. Open-vocabulary video scene graph generation via union-aware semantic alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 8566–8575, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681061. URL https://doi.org/10.1145/3664647.3681061.

Shuo Yang, Yongqi Wang, Xiaofeng Ji, and Xinxiao Wu. Multi-modal prompting for open-vocabulary video visual relationship detection. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i7.28472. URL https://doi.org/10.1609/aaai.v38i7.28472.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, and Manling Li. Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8579–8591, June 2025.

Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 5281–5291, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3612096. URL https://doi.org/10.1145/3581783.3612096.

Yuan Zhou, Litao Hua, Shilong Jin, Wentao Huang, and Haoran Duan. Reason: Reinforced causal search with information bottleneck for video understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13791–13799, 2026.

Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. Air: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. In *International Conference on Learning Representations*, volume 2026, pp. 11302–11329, 2026.

A Detailed Formulation of the Adaptive Evidence Extraction Framework

A.1 Temporal Change Measures

This section provides the detailed definitions of the three temporal change operators used in the boundary score. We follow the notation introduced in the main paper. For a timestamp t, let $\mathcal{T}_t$ denote the trajectories visible at frame t, and define the set of trajectories shared by two consecutive frames as $\mathcal{T}_t^{\cap} = \mathcal{T}_t \cap \mathcal{T}_{t-1}$.

The global motion change is computed from two consecutive global motion representations:

$$\Delta_g(\mathbf{m}^g; t) = \left\| \mathbf{m}^g_{f_t} - \mathbf{m}^g_{f_{t-1}} \right\|_2 . \tag{19}$$

For each trajectory $T^n \in \mathcal{T}_t^{\cap}$, let $\mathbf{m}^n_{f_t}$ denote the motion feature pooled from the image region covered by the corresponding bounding box. The object-region motion change is defined as

$$\Delta_m(\mathbf{m}^g, \mathcal{T}; t) = \begin{cases} \dfrac{1}{|\mathcal{T}_t^{\cap}|} \displaystyle\sum_{T^n \in \mathcal{T}_t^{\cap}} \left\| \mathbf{m}^n_{f_t} - \mathbf{m}^n_{f_{t-1}} \right\|_2, & |\mathcal{T}_t^{\cap}| > 0, \\ 0, & |\mathcal{T}_t^{\cap}| = 0. \end{cases} \tag{20}$$

Let $\mathbf{b}^n_{f_t}$ denote the normalized bounding-box representation of trajectory T^n at frame f_t. The object-layout change is computed as

$$\Delta_l(\mathcal{T}; t) = \begin{cases} \dfrac{1}{|\mathcal{T}_t^{\cap}|} \displaystyle\sum_{T^n \in \mathcal{T}_t^{\cap}} \left\| \mathbf{b}^n_{f_t} - \mathbf{b}^n_{f_{t-1}} \right\|_2, & |\mathcal{T}_t^{\cap}| > 0, \\ 0, & |\mathcal{T}_t^{\cap}| = 0. \end{cases} \tag{21}$$

Each change sequence is independently normalized within a video using min–max normalization:

$$\widehat{\Delta}_q(t) = \frac{\Delta_q(t) - \min_\tau \Delta_q(\tau)}{\max_\tau \Delta_q(\tau) - \min_\tau \Delta_q(\tau) + \varepsilon}, \qquad q \in \{g, m, l\}, \tag{22}$$

where ε is a small positive constant. The resulting normalized measures are substituted into the boundary score defined in the main paper.

A.2 Boundary Detection and Temporal Segmentation

We detect candidate temporal boundaries from the boundary-score sequence $\{D_t\}_{t=2}^{L}$. A timestamp t is considered a local maximum if its score is not smaller than the scores within a temporal neighborhood of radius w_{loc}:

$$D_t \geq D_\tau, \qquad \forall \tau \in [t - w_{\mathrm{loc}}, t + w_{\mathrm{loc}}] \cap \{2, \ldots, L\}. \tag{23}$$

The detected local maxima are sorted in descending order according to their boundary scores. We then apply temporal non-maximum suppression. Specifically, after retaining a candidate boundary at timestamp t, all remaining candidates within a temporal distance smaller than w_{nms} from t are suppressed. This process is repeated until no candidate remains.

The first and last video frames are added as mandatory temporal boundaries. The retained boundary indices are then sorted chronologically and partition the video into contiguous segments:

$$S_j = [\tau_{j-1}, \tau_j), \qquad j = 1, \ldots, M, \tag{24}$$

where $\tau_0 = 1$ and $\tau_M = L + 1$. After boundary selection, segments shorter than $L_{\min} = 10$ frames are merged with an adjacent segment by removing the shared internal boundary. The parameters used for local-maximum detection, temporal non-maximum suppression, and segment length control are set to $w_{\mathrm{loc}} = \mathbf{3}, w_{\mathrm{nms}} = \mathbf{10}, L_{\min} = \mathbf{10}$.

The resulting segments are shared by all candidate trajectory pairs. They serve only as temporal units for adaptive evidence allocation and are not treated as manually annotated relationship boundaries.

A.3 Frame Quality and Diversity Scores

For each valid candidate frame $t \in \mathcal{C}_{p,j}$, we compute a box-separation score $c_{p,j}(t)$ and a frame sharpness score $s_{p,j}(t)$. Let b_t^s and b_t^o denote the subject and object bounding boxes in frame t. We measure the spatial overlap between the two boxes using their intersection-over-union:

$$\mathrm{IoU}(b_t^s, b_t^o) = \frac{\mathrm{area}(b_t^s \cap b_t^o)}{\mathrm{area}(b_t^s \cup b_t^o) + \varepsilon}. \tag{25}$$

The box-separation score is defined as

$$c_{p,j}(t) = 1 - \mathrm{IoU}(b_t^s, b_t^o). \tag{26}$$

A larger value of $c_{p,j}(t)$ indicates less spatial overlap between the subject and object boxes. We measure image sharpness using the variance of the Laplacian response within each trajectory box. Let I_t^x denote the image crop within bounding box b_t^x. The frame sharpness score is defined as

$$s_{p,j}(t) = \frac{1}{2}\left[\mathrm{Var}\left(\nabla^2 I_t^s\right) + \mathrm{Var}\left(\nabla^2 I_t^o\right)\right], \tag{27}$$

where ∇^2 denotes the Laplacian operator. A higher value indicates a sharper crop.

Both scores are normalized over the candidate set $\mathcal{C}_{p,j}$ to make them comparable across frames:

$$\widetilde{y}_{p,j}(t) = \frac{y_{p,j}(t) - \min_{u \in \mathcal{C}_{p,j}} y_{p,j}(u)}{\max_{u \in \mathcal{C}_{p,j}} y_{p,j}(u) - \min_{u \in \mathcal{C}_{p,j}} y_{p,j}(u) + \varepsilon}, \quad y \in \{c, s\}, \tag{28}$$

where ε is a small positive constant.

Visual redundancy is measured by the cosine similarity between ℓ_2-normalized global visual features:

$$\mathrm{sim}\left(\mathbf{v}_{f_t}^g, \mathbf{v}_{f_{t'}}^g\right) = \frac{(\mathbf{v}_{f_t}^g)^\top \mathbf{v}_{f_{t'}}^g}{\|\mathbf{v}_{f_t}^g\|_2 \|\mathbf{v}_{f_{t'}}^g\|_2}. \tag{29}$$

Given a set of already selected frames $\mathcal{A}$, the quality score of a candidate frame is defined as

$$Q_{p,j}(t \mid \mathcal{A}) = \eta_c \widetilde{c}_{p,j}(t) + \eta_s \widetilde{s}_{p,j}(t) - \lambda_{\mathrm{red}} \max_{t' \in \mathcal{A}} \mathrm{sim}\left(\mathbf{v}_{f_t}^g, \mathbf{v}_{f_{t'}}^g\right), \tag{30}$$

where the final term penalizes visual redundancy with the already selected frames. When $\mathcal{A} = \emptyset$, the redundancy term is set to zero. The weights η_c, η_s, and λ_{red} are set according to the values reported in the main paper. For simplicity, we use $c_{p,j}(t)$ and $s_{p,j}(t)$ to denote their normalized values in the frame-selection score.

A.4 Greedy Frame Selection

For each valid pair–segment instance (p, j), we initialize the selected set as $\mathcal{A}_0 = \emptyset$. At each iteration, we select the remaining candidate frame with the highest quality score:

$$t_r = \arg \max_{t \in \mathcal{C}_{p,j} \setminus \mathcal{A}_{r-1}} Q_{p,j}(t \mid \mathcal{A}_{r-1}), \tag{31}$$

and update the selected set as

$$\mathcal{A}_r = \mathcal{A}_{r-1} \cup \{t_r\}. \tag{32}$$

The score is recomputed with respect to the updated selected set at each iteration, so that the redundancy penalty reflects the frames selected so far. The procedure terminates when $|\mathcal{A}_r| = \widehat{k}_{p,j}$ or when all candidate frames have been selected. The resulting set is denoted by

$$\mathcal{F}_{p,j} = \mathcal{A}_r.$$

Finally, the selected frames are arranged in chronological order before being passed to the downstream relationship classifier. Since $\widehat{k}_{p,j}$ is clipped by $|\mathcal{C}_{p,j}|$, all selected frames belong to the candidate set and no frame is selected more than once.

A.5 Implementation of Evidence Extraction Modules

Observation Requirement Predictor. The observation requirement predictor is implemented as a regression network. It takes the pair-specific appearance and motion features as input and predicts a continuous observation requirement:

$$\widetilde{k}_{p,j} = g_{\text{reg}}\left(\left[\mathbf{v}^c_{p,j}; \mathbf{m}^c_{p,j}\right]\right), \qquad \widetilde{k}_{p,j} \in \mathbb{R}. \tag{33}$$

The continuous prediction is rounded and clipped to obtain a feasible integer observation count:

$$\widehat{k}_{p,j} = \text{clip}\left(\text{round}(\widetilde{k}_{p,j}), 1, \min(K_{\max}, |\mathcal{C}_{p,j}|)\right). \tag{34}$$

Optical-Flow Encoder. For each pair–segment instance, the optical-flow encoder takes the pair-specific global flow representation $\mathbf{m}^g_{p,j}$ as input and produces a motion feature:

$$\mathbf{m}^c_{p,j} = E_{\text{flow}}\left(\mathbf{m}^g_{p,j}\right). \tag{35}$$

The input flow representation is separated into the subject, object, and union regions at different spatial resolutions. Each region is processed by a convolutional block consisting of stacked convolutional layers, nonlinear activation functions, and adaptive average pooling. The resulting regional features are concatenated and projected to obtain the final pair-specific motion representation:

$$\mathbf{h}^{\text{flow}}_{p,j} = \text{Concat}\left(\left\{\text{Pool}\left(\text{ConvBlock}\left(\mathbf{m}^{g,r,q}_{p,j}\right)\right)\right\}_{r,q}\right), \tag{36}$$

$$\mathbf{m}^c_{p,j} = P_{\text{flow}}\left(\mathbf{h}^{\text{flow}}_{p,j}\right), \tag{37}$$

where $r \in \{s, o, u\}$ denotes the subject, object, and union regions, and q indexes the spatial resolutions.

Box-Guided Feature Extraction. The box-guided feature extractor takes the pair-specific visual representation $\mathbf{v}^g_{p,j}$ and the corresponding subject and object bounding boxes as input:

$$\mathbf{v}^c_{p,j} = E_{\text{box}}\left(\mathbf{v}^g_{p,j}, T^s, T^o\right). \tag{38}$$

Specifically, box-guided local visual features are first extracted from the subject–object regions over the temporal dimension:

$$\mathbf{z}_{p,j,t} = E^{\text{local}}_{\text{box}}\left(\mathbf{v}^g_{p,j,t}, b^s_t, b^o_t\right), \qquad t \in \mathcal{C}_{p,j}. \tag{39}$$

The resulting feature sequence is encoded by a bidirectional recurrent network and aggregated with attention pooling:

$$\mathbf{h}_{p,j,1:|\mathcal{C}_{p,j}|} = E_{\text{BiGRU}}\left(\{\mathbf{z}_{p,j,t}\}_{t\in\mathcal{C}_{p,j}}\right), \tag{40}$$

$$\mathbf{v}^c_{p,j} = P_{\text{box}}\left(\text{AttnPool}\left(\mathbf{h}_{p,j,1:|\mathcal{C}_{p,j}|}\right)\right). \tag{41}$$

The temporal validity mask is used to exclude invalid or padded features during recurrent encoding and attention pooling.

B Observation Requirement Supervision

B.1 MLLM Prompt and Input Construction

For each base relationship category r in each dataset, we randomly select $N_{\text{clip}} = 20$ annotated relationship instances from the corresponding training split using a fixed random seed. No validation or test clips are used to construct the relationship-level temporal prior. For each selected instance, we extract the video interval corresponding to its annotated temporal extent and process the resulting clip at 30 FPS. The subject and object bounding boxes are rendered in red and blue, respectively. The resulting 20 clips are jointly provided to the MLLM, together with the name of the relationship category. Apart from the relationship name and the annotations used solely to extract the temporal interval and visualize the subject–object trajectories, no object category labels, numerical bounding-box coordinates, captions, or other semantic annotations are provided to the MLLM.

We use the following fixed prompt:

```
You are given 20 video clips at 30 FPS showing the relationship
[RELATIONSHIP]. The subject is indicated by a red bounding box,
and the object is indicated by a blue bounding box.

Analyze all clips jointly and estimate a typical number of sampled
visual frames that a relationship recognition model should retain
from one relationship interval to recognize this relationship.
The frames are observations sampled across the interval and do not
need to be consecutive.

Return one integer in the range [1, 24].
Return the integer only, without any explanation.
```

Here, "[RELATIONSHIP]" is replaced with the name of the corresponding base relationship category. The returned integer is treated as a coarse relationship-level weak prior rather than the ground-truth optimal frame count. During pseudo-target construction, this prior is further adapted to each relationship instance and clipped according to the available candidate-frame capacity. We use the `gpt-5` model with low reasoning effort, low output verbosity, and a maximum output length of 8 tokens. The model is prompted to return a single integer in the range $[1, 24]$. Outputs are parsed as integers and clipped to this range when necessary. The resulting priors are generated once offline and fixed for all experiments. The MLLM is not invoked during model optimization or inference.

B.2 Capacity-Constrained Target Allocation

The segment-level targets are obtained using a capacity-constrained largest-remainder procedure. For each relationship instance ρ, we first assign one observation to every valid segment in $\mathcal{S}_\rho^+$. We then distribute the remaining observation budget according to the allocation weights $\{w_{\rho,j}\}$.

Specifically, let B_ρ denote the remaining budget after the initial assignment. The ideal fractional allocation for each segment is computed as

$$a_{\rho,j} = B_\rho w_{\rho,j}. \tag{42}$$

We first assign $\min\left(\lfloor a_{\rho,j} \rfloor, K_{\rho,j}^{\mathrm{cap}} - 1\right)$ additional observations to each segment. If some observations remain unassigned, they are allocated one at a time according to the fractional remainders $a_{\rho,j} - \lfloor a_{\rho,j} \rfloor$. At each step, the segment with the largest remaining fractional remainder is selected, provided that its current target is below $K_{\rho,j}^{\mathrm{cap}}$. Ties are broken by the allocation weight $w_{\rho,j}$ and then by temporal order. The procedure terminates when the complete budget has been assigned or all segments reach their capacities.

When multiple relationship instances share the same trajectory-pair and segment, their allocated targets are handled according to the aggregation rule defined in the main paper. The resulting segment-level targets are used as weak supervision for the observation requirement predictor.

C Implementation Details

The downstream object classifier is retrained based on the trajectory classifier of EOV-MMP Wang et al. (2025a). In particular, the classifier is trained using the frames selected by the adaptive evidence extraction module rather than using all frames from the original video. The trajectory-level object classification scores provided by the preprocessing stage are retained as trajectory-level evidence, while the object classifier is optimized on the sampled frame representations. The trajectory-level and sampled-frame object predictions are fused according to the formulation described in the main paper.

For relationship classification, we follow the relationship recognition architecture of EOV-MMP Wang et al. (2025a). The selected frames, pair-specific appearance features, and pair-specific motion features are provided to the visual, geometric, motion, and optical-flow branches of the relationship classifier. The resulting representations are aggregated in the temporal order of the selected frames and used to predict the multi-label relationship scores. The parameters of the downstream object and relationship classifiers are optimized using the sampled evidence generated by the proposed frame selection module.

Table 8: Ablation of temporal segmentation strategies on VidVRD .

Segmentation strategy	Novel mAP	All mAP	Avg. Segments↓
Fixed-length segmentation	16.45	25.22	5.27
Global motion only	17.05	26.88	3.64
Object-region motion only	17.16	27.10	3.48
Object-layout change only	17.33	26.97	3.18
Full motion-aware segmentation	**18.08**	**28.93**	**2.79**

Table 9: Ablation of the maximum sampling budget $K_{\max}$ on VidVRD.

$K_{\max}$	Novel mAP	All mAP	Avg. Frames↓
8	16.93	27.04	**2.53**
16	17.31	27.85	2.85
24	**18.08**	**28.93**	3.12
32	17.89	28.37	3.30

C.1 Evidence Extraction Modules

The hidden dimension of the evidence extraction modules is set to 768. The optical-flow encoder contains six independent convolutional branches, corresponding to the subject, object, and union regions at 7×7 and 14×14 spatial resolutions. Each branch consists of two convolutional layers with 32 and 64 output channels, respectively, followed by adaptive average pooling to a 2×2 spatial resolution. The resulting features are concatenated and projected to a 768-dimensional motion representation. The box-guided feature extractor uses a bidirectional GRU with hidden dimension 384 in each direction. Its bidirectional output is therefore 768-dimensional. The attention-pooling module and the output projection layer preserve the final feature dimension of 768. The observation requirement predictor is a regression network with a 768-dimensional hidden representation and a one-dimensional scalar output.

C.2 Training Configuration

We adopt a two-stage training strategy. For the first stage, we use the AdamW optimizer with an initial learning rate of 0.001, a weight decay of 0.0001, and a batch size of 32. The evidence extraction modules are trained for 5 epochs. For the second stage, we use the AdamW optimizer with an initial learning rate of 0.0001, a weight decay of 0.001, and a batch size of 32. The model is trained for 40 epochs. The coefficient of the observation requirement loss is set to $\lambda_{\mathrm{c}} = 0.5$.

During the second stage, the straight-through cumulative mask uses a temperature $\gamma = 1.0$ and an offset $\delta = 0.5$. The maximum number of selected observations is set to $K_{\max} = 24$. Unless otherwise specified, all other optimization and training settings follow EOV-MMP Wang et al. (2025a).

D Additional Experiments

D.1 Segmentation Ablation

We evaluate the contribution of the proposed motion-aware temporal segmentation module on the VidVRD test set under the SGDet setting. All other components, including observation-count prediction, frame selection, and relationship classification, are kept unchanged. We compare fixed-length segmentation with segmentation based on different combinations of global motion, object-region motion, and object-layout changes. Table 8 reports relationship mAP, the average number of temporal segments per trajectory pair. The results show how the temporal segmentation strategy affects both recognition performance and the amount of visual evidence used by the model.

D.2 Sampling Budget Ablation

We study the effect of the maximum number of frames allowed for each trajectory-pair–segment instance. We evaluate $K_{\max} \in \{8, 16, 24, 32\}$ under the SGDet setting, while keeping the model architecture and all other hyperparameters unchanged. To avoid selecting $K_{\max}$ on the test set, we randomly hold out 10% of the original training data as a validation set and select $K_{\max}$ based exclusively on its performance on this validation split. For each setting, the capacity constraints used in pseudo-target construction are adjusted according to the corresponding value of $K_{\max}$. After hyperparameter selection, the model with the selected $K_{\max}$ is evaluated once on the VidVRD test set.

We report Novel mAP, All mAP, and the average number of selected frames per trajectory pair for the validation experiments. These results illustrate the trade-off between recognition performance and the amount of visual evidence used by the model, without using the test set for hyperparameter selection.

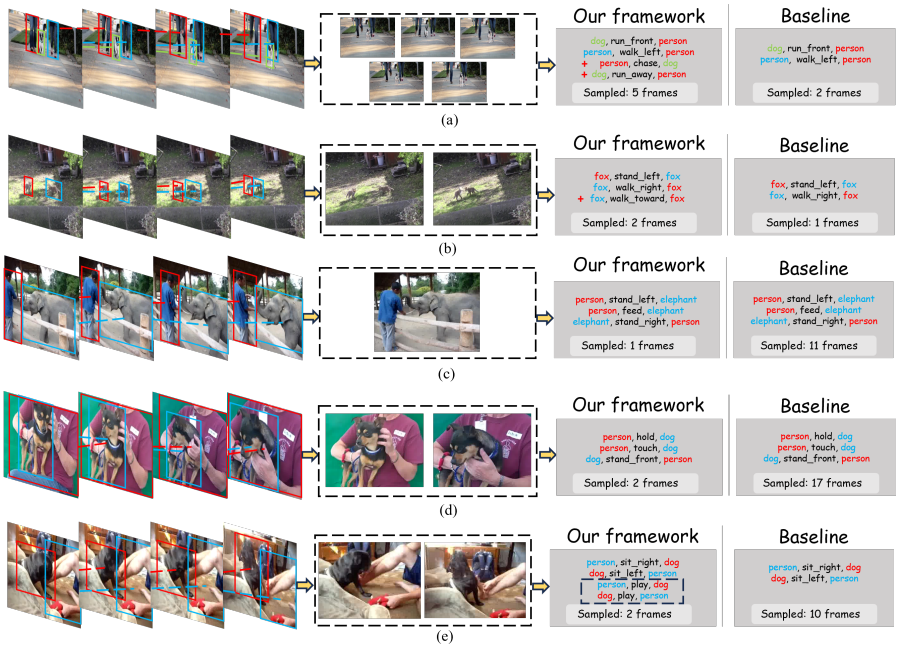

Figure 4: Qualitative case studies comparing our framework with baseline. The red "+" symbols highlight additional relationship triplets correctly predicted by our framework but missed by the baseline, while the black dashed boxes indicate relationship triplets that are not correctly predicted by our framework.

E Additional Qualitative Results and Failure Cases

Figure 4 presents qualitative comparisons between the baseline and our adaptive visual evidence extraction framework. Panels (a) and (b) contain videos with substantial changes in object motion and spatial layout. In these examples, our framework selects more frames than the baseline to capture the changing visual evidence observed in the video. The red "+" symbols indicate additional relationship triplets that are correctly predicted by our framework but missed by the baseline. Panels (c) and (d) contain videos with relatively stable object appearance and spatial configurations. In these cases, our framework selects substantially fewer frames while preserving the relevant relationship predictions. These examples show that our framework adjusts the amount of visual evidence according to the observed visual dynamics, rather than uniformly increasing or reducing the sampling budget. Panel (e) presents a representative failure case in which the subject and object exhibit limited observable motion. As a result, the sampled frames do not provide sufficiently discriminative evidence for recognizing the fine-grained action relationship, and our framework fails to predict the corresponding relationship triplets, as indicated by the black dashed boxes. This case shows that motion-aware frame allocation may remain insufficient when subtle action differences are not clearly expressed through visible motion.