PACT: PAIR-AWARE CONTEXTUAL TOKENS FOR VIDEO VISUAL RELATION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video Visual Relation Detection (VidVRD) focuses on detecting ⟨subject, predicate, object⟩ relations together with their temporal intervals in videos. Most existing methods construct relation representations from features extracted from detected object regions. Consequently, contextual information from the surrounding scene is largely underutilized, and the representations remain sensitive to detection and tracking quality. However, many predicates depend on how two objects interact within the broader scene. We propose Pair-Aware Contextual Tokens (PACT), a relation representation that combines full-frame scene context with information about the target subject–object pair. PACT uses full-frame visual tokens to capture context beyond the subject and object regions and directly injects pair-specific spatial information and category semantics into these tokens. Experiments on `ImageNet-VidVRD` and `VidOR` show that PACT achieves state-of-the-art mAP on both benchmarks. These results demonstrate that preserving global scene context while incorporating pair-specific spatial and semantic information provides an effective relation representation for VidVRD.

1 INTRODUCTION

Video Visual Relation Detection (VidVRD) aims to identify how objects interact with each other over time in a video. Each relation is represented as a triplet ⟨subject, predicate, object⟩, together with the temporal interval during which the relation holds. By providing a compact, interpretable description of object interactions, VidVRD supports various downstream tasks, including video captioning, video question answering, and video retrieval. Unlike image relation detection, VidVRD must capture not only the spatial interaction between the subject and object in each frame but also how this interaction evolves across frames. Object appearances vary due to motion and occlusion, and each relation is valid only over a limited temporal interval. Learning a relation representation that reflects the surrounding scene context and temporal dynamics is therefore a central challenge in VidVRD.

Existing VidVRD methods (Shang et al., 2017; 2021; Zheng et al., 2022; Cong et al., 2023) infer relations from entity tubelets generated through object detection and tracking. Most methods construct relation representations from features localized to the paired entities, using region pooling or entity-centric queries. Recent studies have improved relation prediction through more effective relation modeling built on these representations. However, the underlying representations often remain object-centric. We argue that such object-localized representations limit relation prediction because they capture limited information about the scene beyond the two entity regions.

To address these limitations, we propose **P**air-**A**ware **C**ontextual **T**okens (PACT), a relation representation that preserves global scene context while incorporating information about the target subject–object pair. Instead of constructing relation representations from object regions, PACT uses the full set of frame-level visual tokens extracted by the image encoder, retaining fine-grained visual details across the entire scene. The intuitive difference from existing relation representations is illustrated in Fig. 2. The spatial information and category semantics of the target subject–object pair are then injected into the visual tokens. This allows the representation to jointly capture global scene context, pair-specific spatial information, and entity semantics.

Building on PACT, we develop a relation prediction framework that models interactions across space and time. The pair-aware contextual tokens are processed by a spatio-temporal transformer to

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.

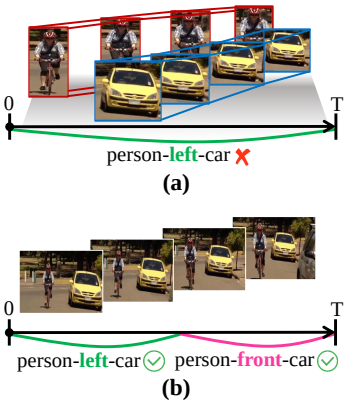

Figure 1: **Importance of scene context.** (a) Object regions alone lack predicate cues. (b) Surrounding scenes enable reliable prediction.

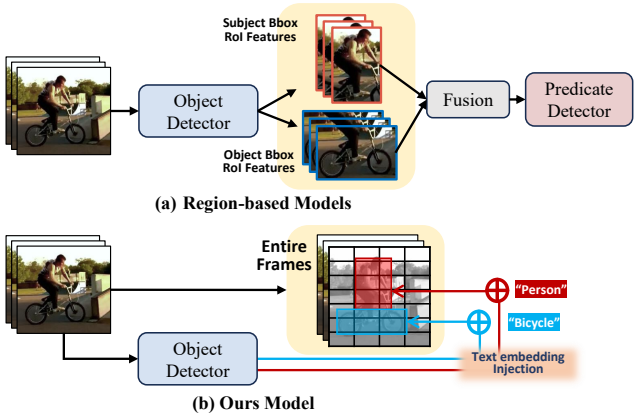

Figure 2: **Comparison of relation representations.** (a) Existing methods use only subject and object regions. (b) PACT uses full-frame visual tokens and specializes the target pair by injecting category embeddings at their regions.

capture spatial interactions within each frame and temporal dynamics across the video. A relation decoder then predicts the predicate and temporal interval for each pair. We evaluate PACT on the `ImageNet-VidVRD` (Shang et al., 2017) and `VidOR` (Shang et al., 2019) benchmarks. Experimental results show that PACT achieves state-of-the-art mAP on both benchmarks and consistently improves relation prediction over object-localized baselines. These results demonstrate the importance of combining global scene context with pair-specific spatial and semantic information for relation prediction.

The main contributions of this work are as follows:

- We identify a fundamental limitation of object-localized relation representations: they underutilize surrounding scene context and remain sensitive to detection and tracking quality.
- We propose Pair-Aware Contextual Tokens (PACT), which specialize full-frame visual tokens for each subject–object pair by incorporating pair-specific spatial information and category semantics.
- PACT consistently improves both Recall and Precision on the `ImageNet-VidVRD` and `VidOR` benchmarks, achieving state-of-the-art mAP on both.

To our knowledge, it is the first VidVRD method to construct relation representations directly from full-frame visual tokens.

2 Related Works

Video Visual Relation Detection (VidVRD). VidVRD was first formulated as a multi-stage pipeline (Shang et al., 2017), where videos are split into short clips, relations are classified within each clip, and the results are associated into longer intervals. Subsequent studies improved this pipeline by modeling trajectory interactions with graph networks (Qian et al., 2019), integrating multi-modal features (Su et al., 2020), or enhancing spatial and temporal context modeling (Liu et al., 2020; Wei et al., 2024).

More recent methods have moved toward single-stage designs. VRDFormer (Zheng et al., 2022) predicts visual relations end-to-end using learnable relation queries in a DETR-style framework (Carion et al., 2020), whereas VrdONE (Jiang et al., 2024) reformulates relation classification and temporal localization as a 1D temporal instance segmentation problem. However, these methods mainly focus on relation reasoning after constructing the initial representation, leaving the design of context-preserving pair representations relatively underexplored.

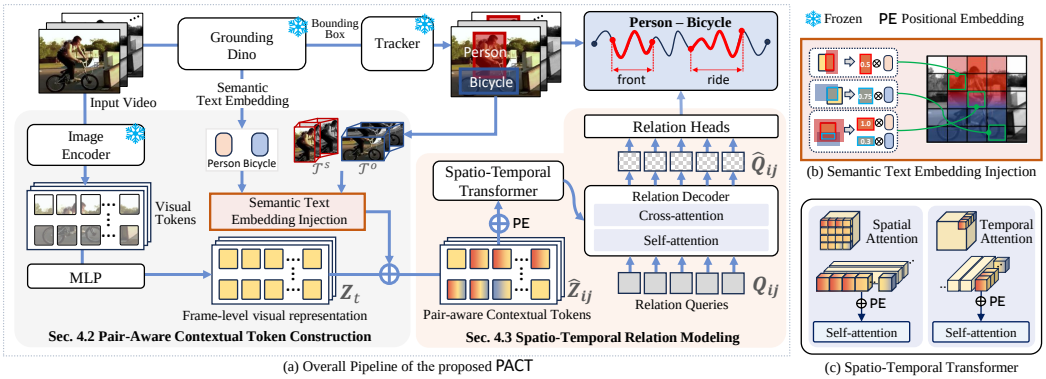

Figure 3: **Overview of PACT.** Given an input video, PACT constructs pair-aware contextual token representations by combining visual patch tokens with spatial and semantic information. The enriched tokens are processed by a spatio-temporal transformer to model subject–object interactions across frames, and a relation decoder predicts the predicate label.

Relation Representation. The limitations of region-based features have long been recognized in image-based relation detection, where relation representations are enriched through graph neural networks (Scarselli et al., 2008), linguistic priors (Zellers et al., 2018), or transformer-based designs that aggregate image-wide context (Kim et al., 2021; Tamura et al., 2021). Despite their effectiveness, these methods do not explicitly specialize shared dense visual tokens for individual subject–object pairs, which is crucial for distinguishing multiple concurrent relations within the same scene.

Recent video transformers such as ViViT (Arnab et al., 2021) and VideoMAE (Tong et al., 2022) learn token-based visual representations without explicit object-level pooling. However, relation representations in VidVRD are often constructed from object- or tracklet-centric features, limiting the surrounding scene context available for relation modeling. To address this limitation, we inject the category semantics and spatial overlap information of each subject–object pair into shared frame-level visual tokens, producing pair-aware contextual tokens that retain scene context while encoding pair-specific information.

3 Motivation and Main Idea

Most existing Video Visual Relation Detection (VidVRD) methods (Cong et al., 2023; Jiang et al., 2024; Chang et al., 2025) construct relation representations centered on the paired objects, using either region pooling or object-centric queries. Although these representations effectively capture object-level visual information, their object-centered design may limit the use of contextual cues from the surrounding scene as shown in Fig. 2(a). Moreover, relation representations constructed directly from detected object regions can be sensitive to localization errors, as inaccurate regions may omit relevant visual evidence or include irrelevant content. Reliable relation prediction requires both broader scene context and explicit identification of the target subject–object pair.

To this end, we propose Pair-Aware Contextual Tokens (PACT), which builds relation representations from the full set of visual tokens extracted from each frame, as illustrated in Fig. 2(b). Unlike object-localized features, these full-frame visual tokens retain the global scene context while preserving fine-grained visual details across the frame. However, a single set of visual tokens is extracted from each frame and shared across all entity pairs. As a result, the shared tokens alone do not indicate which entities constitute the target subject–object pair. To make the tokens pair-aware, we encode the semantic identities (i.e., object categories) of the subject and object as text embeddings and inject them into the visual tokens according to their corresponding spatial regions. The resulting tokens therefore preserve the full-frame visual context while explicitly incorporating the semantic and spatial information of the target subject and object.

4 PROPOSED METHODS

As illustrated in Fig. 3, the proposed Pair-Aware Contextual Tokens (PACT) framework consists of two main components: Pair-Aware Contextual Token Construction and Spatio-Temporal Relation Modeling. First, PACT constructs pair-aware contextual tokens by incorporating the spatial and semantic information of each subject–object pair into full-frame visual tokens in Sec. 4.2. It then models spatial interactions and temporal dynamics using a spatio-temporal transformer and predicts relations through a relation decoder in Sec. 4.3.

4.1 PROBLEM FORMULATION

The goal of Video Visual Relation Detection (VidVRD) is to detect all visual relation instances between pairs of entities in a video sequence. Each relation instance consists of a subject–predicate–object triplet, its temporal interval, and the corresponding subject and object tubelets. Given a video of T frames $\{I_t\}_{t=1}^{T}$, we use entity tubelets extracted by a pre-trained object detector and tracker. Each detected entity i is assigned an object category $c_i \in \mathcal{C}$, where $\mathcal{C}$ denotes the set of object categories, and is tracked over a contiguous frame interval $\tau_i \subseteq \{1, \dots, T\}$, referred to as its temporal support. Its tubelet, $\mathcal{T}_i = \{\, b_{i,t} \,\}_{t \in \tau_i}$, is the sequence of bounding boxes occupied by the entity over this interval. Here, $b_{i,t}$ denotes the bounding box of entity i at frame t.

Formally, the prediction target is a set of relation instances $\{\mathcal{G}_k\}_{k=1}^{K}$. The k-th relation instance is defined as

$$\mathcal{G}_k = \big(\langle s_k,\, r_k,\, o_k \rangle,\ \mathcal{T}_k^s,\ \mathcal{T}_k^o,\ [\, t_k^b,\, t_k^e \,] \big), \tag{1}$$

where K denotes the number of relation instances in the video. Here, $s_k, o_k \in \mathcal{C}$ denote the subject and object categories, respectively, where $\mathcal{C}$ is the set of object categories. The predicate label is denoted by $r_k \in \mathcal{P}$, where $\mathcal{P}$ is the set of predicate categories. The frame indices t_k^b and t_k^e define the temporal interval $[\, t_k^b,\, t_k^e \,] \subseteq \{1, \dots, T\}$ during which the relation holds. The tubelets $\mathcal{T}_k^s$ and $\mathcal{T}_k^o$ denote the corresponding subject and object tubelets restricted to this interval. A single pair of entities may yield multiple relation instances with different predicates or temporal intervals. For relation prediction, we form ordered subject–object pairs (i, j), where $i \neq j$. We define their shared temporal support as $\tau_{ij} = \tau_i \cap \tau_j$.

4.2 PAIR-AWARE CONTEXTUAL TOKEN CONSTRUCTION

To preserve both the spatial relationships between entities and the surrounding visual context, we represent each frame using the full set of visual tokens, rather than constructing relation representations from object regions. A frozen DINOv2 image encoder (Oquab et al., 2023) extracts visual tokens from each frame I_t, and the resulting tokens are shared across all entity pairs. Each visual token is linearly projected into an H-dimensional hidden space, followed by layer normalization. This yields the frame-level visual representation $\mathbf{Z}_t = [\mathbf{z}_{t,1}; \dots; \mathbf{z}_{t,M}] \in \mathbb{R}^{M \times H}$, where $\mathbf{z}_{t,m} \in \mathbb{R}^H$ denotes the feature corresponding to the m-th spatial patch. Each token represents a local image region, while the ordered token sequence retains the spatial layout of the frame. The resulting representation provides the shared visual foundation for constructing pair-aware contextual tokens.

Since a single set of visual tokens is extracted from each frame and shared across all entity pairs, the resulting token representations do not explicitly indicate which entities constitute the target subject–object pair. To make the shared visual tokens pair-aware, we introduce semantic text embedding injection, which encodes the subject and object categories as text embeddings and injects them into the corresponding visual tokens based on spatial overlap with the target entities. In this way, the shared visual representation is conditioned on both the semantic identities and spatial locations of the target subject and object.

Semantic Text Embedding Injection. For each frame $t \in \tau_{ij}$, we first compute the spatial overlap of each visual token with the subject and object regions. Specifically, we define $\rho_{t,m}^s$ and $\rho_{t,m}^o$ as the overlap scores between the image region corresponding to the m-th visual token and the subject and object regions, respectively:

$$\rho_{t,m}^s = \frac{|\Omega_m \cap b_{i,t}|}{\min(|\Omega_m|, |b_{i,t}|)}, \quad \rho_{t,m}^o = \frac{|\Omega_m \cap b_{j,t}|}{\min(|\Omega_m|, |b_{j,t}|)}, \tag{2}$$

where Ω_m denotes the image region corresponding to the m-th visual token, and $|\cdot|$ denotes the area of a region. Normalization by the smaller region allows the overlap score to remain informative even when an entity occupies an area smaller than a visual token. The resulting overlap vector $\boldsymbol{\rho}_{t,m} = [\rho^s_{t,m}, \rho^o_{t,m}]$ encodes the spatial association of the m-th visual token with the target subject and object.

We then encode the semantic information of the target entities using their detected category labels. Specifically, the subject and object category labels c_i and c_j, obtained from the object detector, are encoded by a pre-trained text encoder and projected into the H-dimensional hidden space, producing the subject and object text embeddings $\mathbf{v}^s, \mathbf{v}^o \in \mathbb{R}^H$. These embeddings are injected into each visual token according to its overlap scores with the corresponding entity regions, as illustrated in Fig. 3(b).

The final pair-aware visual token is defined as

$$\hat{\mathbf{z}}_{t,m} = \mathrm{LN}\left(\mathbf{z}_{t,m} + \gamma_p \phi(\boldsymbol{\rho}_{t,m}) + \gamma_c(\rho^s_{t,m}\mathbf{v}^s + \rho^o_{t,m}\mathbf{v}^o)\right), \tag{3}$$

where $\phi(\cdot)$ is a learnable mapping to $\mathbb{R}^H$, LN denotes layer normalization, and γ_p and γ_c are learnable scalar gates that control the contributions of the spatial and semantic information, respectively. When a visual token overlaps with both entities, the subject and object embeddings are combined according to their respective overlap scores. As a result, the shared visual token $\mathbf{z}_{t,m}$ is transformed into a pair-aware contextual token $\hat{\mathbf{z}}_{t,m}$ while preserving the full-frame visual context. This injection is performed independently for each subject–object pair, producing a distinct pair-aware contextual representation $\hat{\mathbf{Z}}_{ij}$ for every pair (i, j).

4.3 Spatio-Temporal Relation Modeling

The pair-aware contextual tokens $\hat{\mathbf{Z}}_{ij}$ encode the full-frame visual context together with the spatial and semantic information of the target subject and object. To predict the relation between the target entities, the model must capture both their spatial interactions within each frame and their temporal evolution across frames. We therefore employ a spatio-temporal transformer to jointly model these two aspects. Since the length of the shared temporal support τ_{ij} varies across subject–object pairs, each token sequence is resampled to a fixed length L before being processed by the transformer. The resampled pair-aware tokens are denoted by $\widetilde{\mathbf{Z}}_{ij} \in \mathbb{R}^{L\times M\times H}$. Although resampling produces sequences of equal length, the resulting sequence positions do not preserve the original temporal spacing between frames, which may obscure the temporal dynamics of the target pair. To retain the original temporal structure, we add positional encodings derived from the original frame indices, as illustrated in Fig. 4.

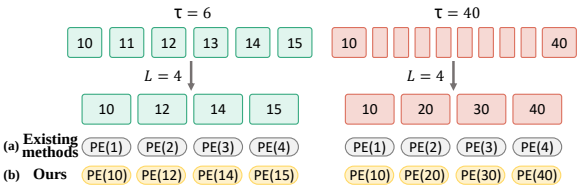

Figure 4: **Positional embedding after temporal resampling.** (a) Existing methods assign positional embeddings from the resampled sequence indices, thereby discarding the original temporal positions of the frames. (b) Our method assigns positional embeddings from the original frame indices, preserving the temporal structure of the video after resampling.

Spatio-Temporal Transformer. The resampled pair-aware contextual representation $\widetilde{\mathbf{Z}}_{ij}$ is processed by a spatio-temporal transformer, as illustrated in Fig. 3(c). Spatial attention is first applied within each frame to model interactions among the contextual tokens. Temporal attention is then applied across time at each spatial index to capture their temporal dynamics:

$$\begin{aligned} \bar{\mathbf{Z}}_{ij}[\ell,:] &= \mathrm{Attn_s}\left(\widetilde{\mathbf{Z}}_{ij}[\ell,:]\right), \quad \ell = 1,\dots,L, \\ \mathbf{F}_{ij}[:,m] &= \mathrm{Attn_t}\left(\bar{\mathbf{Z}}_{ij}[:,m]\right), \quad m = 1,\dots,M, \end{aligned} \tag{4}$$

where $\mathrm{Attn_s}(\cdot)$ and $\mathrm{Attn_t}(\cdot)$ denote spatial and temporal multi-head self-attention, respectively. Here, $\widetilde{\mathbf{Z}}_{ij}[\ell,:] \in \mathbb{R}^{M\times H}$ denotes the set of contextual tokens at the ℓ-th time step, while $\bar{\mathbf{Z}}_{ij}[:,m] \in \mathbb{R}^{L\times H}$ denotes the temporal sequence of the contextual token at the m-th spatial index. The resulting representation $\mathbf{F}_{ij} \in \mathbb{R}^{L\times M\times H}$ captures spatial interactions within each frame and temporal dynamics across frames, and is subsequently used by the relation decoder for predicate prediction.

Relation Decoder and Heads Given the spatio-temporal representation $\mathbf{F}_{ij}$, we adopt a transformer decoder following DETR (Carion et al., 2020) to predict relation instances. The representation $\mathbf{F}_{ij}$ is used as the decoder input, together with N learnable relation queries $\mathbf{Q}_{ij} \in \mathbb{R}^{N\times H}$. Since a single subject–object pair may exhibit multiple predicates over different temporal intervals, we formulate relation prediction as a set prediction problem, where each query represents one relation candidate. The relation queries are refined through self-attention among the queries and cross-attention to $\mathbf{F}_{ij}$. The resulting pair-specific query representations $\hat{\mathbf{Q}}_{ij} \in \mathbb{R}^{N\times H}$ are fed into relation heads that predict the predicate label and temporal span of each relation candidate. Each head is a small feed-forward network, and its detailed configuration is provided in the supplementary material.

4.4 Learning and Inference

Loss Functions. Each relation query is trained to predict a predicate together with the temporal extent of the relation. Since the query outputs are unordered, we first match them to the ground-truth relations using the Hungarian algorithm and then compute the losses based on the resulting assignment. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{span}}\,\mathcal{L}_{\text{span}} + \mathcal{L}_{\text{aux}}, \tag{5}$$

where λ_{span} controls the contribution of the temporal term. $\mathcal{L}_{\text{pred}}$ is the predicate classification loss. Predicate prediction is formulated as a multi-class classification problem with an additional no-relation category and optimized using cross-entropy over all queries. $\mathcal{L}_{\text{span}}$ denotes the temporal localization loss. Instead of directly regressing the temporal boundaries, each matched query predicts frame-level relation confidence scores within the pair's co-occurrence interval τ_{ij}, which are supervised using binary cross-entropy against the ground-truth temporal mask.

$\mathcal{L}_{\text{aux}}$ provides additional supervision for subject and object prediction. It combines an object classification loss $\mathcal{L}_{\text{cls}}$, defined as cross-entropy on the subject and object categories, and a bounding-box regression loss $\mathcal{L}_{\text{box}}$, defined as the ℓ_1 loss on their boxes:

$$\mathcal{L}_{\text{aux}} = \lambda_{\text{cls}}\,\mathcal{L}_{\text{cls}} + \lambda_{\text{box}}\,\mathcal{L}_{\text{box}}, \tag{6}$$

where λ_{cls} and λ_{box} control the contribution of each term. Both auxiliary losses are applied to the matched queries.

Relation Detection. At inference, relation instances are constructed from the decoder queries for each candidate pair (i, j). Each query predicts a predicate distribution together with frame-level relation confidence scores $\hat{\sigma}_t$ over the co-occurrence interval τ_{ij}. The predicate label is selected as the category with the highest predicted probability. The temporal span is recovered from $\hat{\sigma}_t$. Consecutive time steps whose confidence scores exceed a predefined threshold are grouped into an interval, which is then mapped back to the original frame indices to obtain $[t^b, t^e]$. The subject and object tubelets $\mathcal{T}_i^s$ and $\mathcal{T}_j^o$ are temporally cropped to the predicted interval.

Each relation instance with predicate r is assigned the score

$$s(r) = s^{sub} \times s^{obj} \times p(r) \times \bar{\sigma}, \tag{7}$$

where s^{sub} and s^{obj} denote the confidence scores of the subject and object tubelets, respectively, $p(r)$ is the predicted probability of predicate r, and $\bar{\sigma}$ is the average frame-level relation confidence over the predicted temporal span. Finally, queries classified as no-relation are discarded, non-maximum suppression is applied to remove redundant relation instances, and the remaining instances are ranked according to $s(r)$.

5 Experimental Results

5.1 Experimental Settings

Datasets. We conduct experiments on two widely used VidVRD benchmarks, `ImageNet-VidVRD` (Shang et al., 2017) and `VidOR` (Shang et al., 2019). `ImageNet-VidVRD` contains approximately 3 hours of video, including 800 training videos and 200 test videos. It includes 35 object classes and 132 predicate classes. `VidOR` is a larger-scale dataset containing

Table 1: **Performance comparison with existing methods on two VidVRD benchmarks.** Bold and underlined values indicate the best and second-best results for each evaluation metric.

Methods	Relation Detection (RelDet)			Relation Tagging (RelTag)	
	mAP	R@50	R@100	P@1	P@5
Dataset: `ImageNet-VidVRD`					
VIDVRD (Shang et al., 2017)	5.48	6.73	9.06	40.62	29.48
VIDVRD II (Shang et al., 2021)	27.39	15.47	18.81	73.59	57.91
VRD-STGC (Liu et al., 2020)	15.79	12.73	14.29	59.72	42.89
VRDFormer (Zheng et al., 2022)	28.16	17.09	20.43	75.11	<u>58.47</u>
RelTR-v2 (Cong et al., 2023)	<u>30.08</u>	<u>20.14</u>	20.88	**77.62**	57.93
VSTRD (Chang et al., 2025)	30.42	**20.55**	<u>22.23</u>	76.24	**59.63**
Ours	**32.11**	<u>20.14</u>	**23.60**	<u>77.50</u>	57.40
Dataset: `VidOR`					
3DRN (Cao et al., 2021)	2.35	2.73	2.96	54.79	45.33
VIDVRD II (Shang et al., 2021)	9.62	9.46	11.48	59.31	47.92
VRD-STGC (Liu et al., 2020)	6.84	8.02	10.37	46.75	39.21
VRDFormer (Zheng et al., 2022)	9.79	9.85	<u>13.10</u>	61.09	49.19
RelTR-v2 (Cong et al., 2023)	10.52	10.07	12.82	60.42	49.26
VSTRD (Chang et al., 2025)	<u>10.73</u>	<u>11.04</u>	**13.86**	<u>61.12</u>	<u>50.73</u>
Ours	**10.87**	**11.05**	12.96	**61.41**	**51.24**

approximately 98.6 hours of video, with 7,000 training videos, 835 validation videos, and 2,165 test videos. It includes 80 object classes and 50 predicate classes, along with dense annotations of object bounding boxes and temporal relation intervals. Following the standard evaluation protocols, we train the model on the training set while reporting results on the validation set.

Evaluation Metrics. We evaluate our method under two standard settings: Relation Detection (RelDet) and Relation Tagging (RelTag). RelDet evaluates the localization and classification of relation instances, including their subject and object trajectories and temporal extents. A predicted relation instance is considered correct if its subject–predicate–object category triplet matches the ground truth and both the predicted subject and object trajectories achieve a volumetric IoU (vIoU) greater than 0.5 with their corresponding ground-truth trajectories. We report mean Average Precision (mAP) and Recall@K (R@50 and R@100). RelTag evaluates whether a relation triplet occurs in a video without considering the localization of object trajectories or temporal relation intervals. We report Precision@K (P@1 and P@5).

Implementation Details. We use frozen Grounding DINO (Liu et al., 2024) to obtain object tubelets and to encode the object categories as semantic text embeddings, and frozen DINOv2 (Oquab et al., 2023) as the visual backbone. Both the spatial and temporal encoders contain two transformer blocks with four attention heads, followed by a three-layer decoder with 10 queries. We train the model using AdamW with a learning rate of 3×10^{-5}, a weight decay of 10^{-4}, and a batch size of one video per GPU. The learning rate is reduced by a factor of 0.1 after 60 epochs, with gradient clipping at 1.0 and dropout set to 0.3. The training objective combines Hungarian-matched predicate classification, auxiliary object classification, and box regression losses, with $\lambda_{\text{cls}} = 0.7$ and $\lambda_{\text{box}} = 2.0$. All models are trained on eight NVIDIA H200 GPUs. Additional implementation details are provided in the appendix.

5.2 Performance Comparison

We compare the proposed PACT against state-of-the-art VidVRD methods on the `ImageNet-VidVRD` (Shang et al., 2017) and `VidOR` (Shang et al., 2019) benchmarks. As shown in Tab. 1, PACT achieves the highest mAP on both benchmarks, with scores of 32.11 on `ImageNet-VidVRD` and 10.87 on `VidOR`. Since mAP reflects both precision–recall performance and the ranking of relation predictions, these results indicate that the proposed pair-aware contextual representation enables more reliable relation scoring.

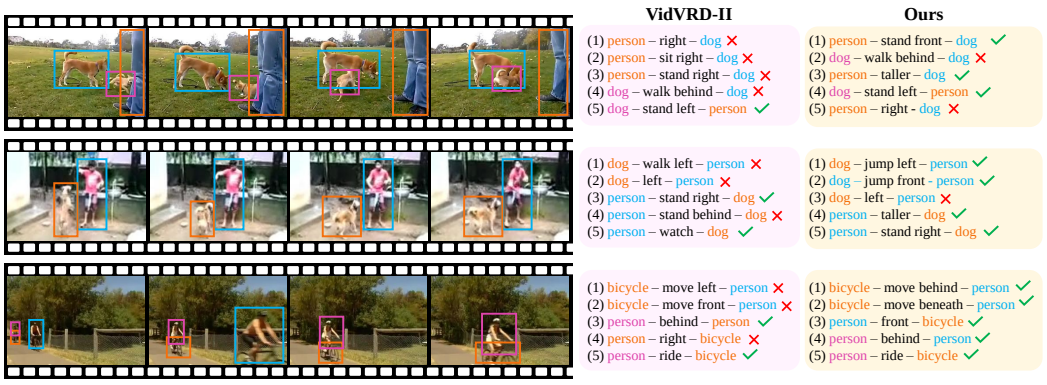

Figure 5: **Qualitative comparison of the top-5 relation predictions from VidVRD-II (Shang et al., 2021) and PACT on `ImageNet-VidVRD`.** Each object is represented by the same color in its bounding box and the corresponding relation triplets. Check marks and crosses indicate correct and incorrect predictions, respectively. More results are provided in the supplementary.

The improvement is most pronounced under the RelDet setting. On `ImageNet-VidVRD`, PACT surpasses VSTRD (Chang et al., 2025) by 1.69 mAP and achieves the highest R@100 of 23.60, suggesting that scene-level context helps capture relations that may be missed by region-based representations. On `VidOR`, PACT likewise achieves the highest mAP and R@50. Unlike RelDet, RelTag evaluates relation triplet classification at the video level without considering trajectory or temporal localization. PACT achieves the highest P@1 and P@5 on `VidOR`. On `ImageNet-VidVRD`, it achieves a P@1 comparable to the best-performing method, although its P@5 is lower. These results indicate that PACT reliably identifies the most likely relation triplet, while its performance decreases when a broader set of predictions is considered.

5.3 Predicate and Qualitative Analysis

To examine which predicates benefit from the proposed representation, we compare per-predicate Recall@100 on `ImageNet-VidVRD` in Fig. 6. Our method improves recall for all ten of the most frequent predicates. Many of these predicates, such as "*stand_left*", "*walk_behind*", and "*taller*", depend primarily on the relative spatial arrangement or motion of the two objects rather than on their individual appearances. Predicting such predicates requires scene-level visual evidence beyond the object regions, to which object-localized representations provide only limited access.

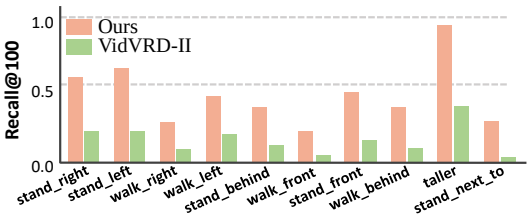

Figure 6: **Per-predicate recall@100 for the ten most frequent predicates on `ImageNet-VidVRD`.** PACT (Ours) consistently outperforms VidVRD-II (Shang et al., 2021) across all ten predicates.

Figure 5 further illustrates this advantage through qualitative examples. Compared with VidVRD-II (Shang et al., 2021), our method ranks more correct relation predictions among the top results. In the second row, our method predicts "*dog–jump left–person*", whereas the baseline predicts "*dog–walk left–person*", assigning an incorrect action to the same object pair. In the third row, all five predictions from our method are correct, while the baseline incorrectly predicts the motion directions in "*bicycle–move left–person*" and "*bicycle–move front–person*". These examples involve either the relative arrangement of the two objects or their motion within the scene, both of which are difficult to determine from object appearance alone.

5.4 Ablation Study

We analyze the contribution of each component of the proposed method. All ablation studies are conducted on `ImageNet-VidVRD` under the same setting as the main experiments. Additional experiments are provided in the supplementary materials.

Table 2: **Ablation study on visual representation design**

Method	Relation Detection		Relation Tagging	
	mAP	R@50	P@1	P@5
Full-frame w/o Inj	25.23	14.81	65.00	48.75
Object-region only	31.68	19.81	77.00	55.90
(**Ours**) Full-frame w Inj	**32.11**	**20.14**	**77.50**	**57.40**

Table 2 compares three visual representations. The first uses full-frame visual tokens without semantic text embedding injection (Full-frame w/o Inj). The second retains only the visual tokens that overlap the subject or object regions (Object-region only). The third uses our pair-aware contextual tokens (Full-frame w Inj). Full-frame visual tokens without injection achieve 25.23 mAP. Although these tokens preserve scene-level context, they are shared across all pairs in a frame and therefore do not explicitly identify the target pair. Retaining only the visual tokens (Object-region only) that overlap the subject or object regions improves mAP to 31.68. This result indicates that explicitly focusing the representation on the target pair is important for relation prediction. However, this setting excludes most visual information outside the object regions, limiting access to broader scene context. Moreover, because the token selection is based on token-region overlap, tokens that partially overlap the object regions may still retain some surrounding information. Our pair-aware contextual tokens (Ours) further improve mAP to 32.11 and achieve the best performance across all metrics. These results demonstrate the benefit of preserving full-frame context while incorporating pair-specific semantic information.

Table 3: **Ablation study on individual model components**

Configuration	Relation Detection		Relation Tagging	
	mAP	R@50	P@1	P@5
w/o $\mathrm{Attn_s}$	31.30	18.59	74.00	55.43
w/o $\mathrm{Attn_t}$	28.86	18.12	70.00	52.68
w/o ST Transformer	28.47	17.64	72.50	51.85
w/o $\mathcal{L}_{\mathrm{cls}}$	30.14	19.05	70.50	53.78
w/o $\mathcal{L}_{\mathrm{box}}$	29.18	17.44	67.50	52.73
(**Ours**) with all	**32.11**	**20.14**	**77.50**	**57.40**

The contribution of each model component is summarized in Tab. 3. Removing $\mathrm{Attn_t}$ (w/o $\mathrm{Attn_t}$) causes a larger performance drop than removing $\mathrm{Attn_s}$ (w/o $\mathrm{Attn_s}$), indicating that temporal modeling plays a particularly important role in this task. Removing the entire spatio-temporal transformer (w/o ST Transformer) yields the lowest mAP among all configurations, demonstrating that spatial and temporal attention provide complementary information. The auxiliary losses supervise the relation queries with object-level category and spatial information. Removing $\mathcal{L}_{\mathrm{cls}}$ (w/o $\mathcal{L}_{\mathrm{cls}}$) primarily degrades relation tagging performance, suggesting that object classification supervision helps the queries encode object-category information. Removing $\mathcal{L}_{\mathrm{box}}$ (w/o $\mathcal{L}_{\mathrm{box}}$) yields the lowest R@50 and P@1, indicating that spatial supervision helps the queries focus on the relevant object regions.

6 Conclusion and Future Work

We proposed Pair-Aware Contextual Tokens (PACT), a relation representation framework for VidVRD constructed from the full-frame visual tokens of a frozen image encoder. Unlike object-localized representations, which underutilize surrounding scene context and remain sensitive to localization errors, PACT directly injects pair-specific spatial information and category semantics into the visual tokens. The resulting pair-aware contextual tokens retain global scene context while encoding the target subject–object pair and are subsequently processed by a spatio-temporal transformer and relation decoder. Experiments on `ImageNet-VidVRD` and `VidOR` demonstrate that PACT achieves the best mAP on both benchmarks without using a video backbone or fine-tuning the visual encoder. Our ablation studies further show that full-frame context is most effective when the visual tokens are explicitly specialized for the target pair.

Nevertheless, PACT still relies on detected and tracked entity tubelets, making its performance susceptible to missed detections and inaccurate trajectories. Moreover, processing full-frame tokens for multiple entity pairs may increase computational cost in crowded videos. Future work could improve robustness through joint entity and relation modeling, develop more efficient pair-specific token selection, and extend PACT toward open-vocabulary VidVRD by exploiting text-based category representations.

AI USE STATEMENT

Generative AI tools were employed exclusively for manuscript polishing, including grammatical correction, readability enhancement, and LaTeX formatting adjustments. These tools were strictly prohibited from contributing to the core research components, such as theoretical formulation, experimental design, algorithmic implementation, and data analysis. All AI-refined text was thoroughly verified by the authors to preserve the original scientific meaning. The authors assume full responsibility for the entirety of this work.

REPRODUCIBILITY STATEMENT

To facilitate full reproducibility, we provide comprehensive details of the proposed Pair-Aware Contextual Tokens (PACT) framework. The architectural design, including token construction (Sec. 4.2) and spatio-temporal modeling (Sec. 4.3), is explicitly formulated in Section 4. Section 5.1 outlines the experimental setup, covering the datasets (`ImageNet-VidVRD` and `VidOR`), evaluation protocols, and precise implementation parameters, such as optimizer settings and the use of frozen models. Extended network configurations are included in the supplementary material. The complete source code and pretrained model weights will be made publicly accessible upon publication.

REFERENCES

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6836–6846, 2021.

Qianwen Cao, Heyan Huang, Xindi Shang, Boran Wang, and Tat-Seng Chua. 3-d relation network for visual relation recognition in videos. *Neurocomputing*, 432:91–100, 2021.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. URL `https://arxiv.org/abs/2005.12872`.

Ruyi Chang, Haopeng Wang, Zhitian Zhang, Dejiao Huang, and Shuai Guo. Vstrd: An end-to-end video spatio-temporal relation detection transformer. *Neurocomputing*, pp. 131307, 2025.

Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. Reltr: Relation transformer for scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11169–11183, 2023.

Xinjie Jiang, Chenxi Zheng, Xuemiao Xu, Bangzhen Liu, Weiying Zheng, Huaidong Zhang, and Shengfeng He. Vrdone: One-stage video visual relation detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1437–1446, 2024.

Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 74–83, 2021.

Chenchen Liu, Yang Jin, Kehan Xu, Guoqiang Gong, and Yadong Mu. Beyond short-term snippet: Video relation detection with spatio-temporal global context. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10840–10849, 2020.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. Video relation detection with spatio-temporal graph. In *Proceedings of the 27th ACM international conference on multimedia*, pp. 84–93, 2019.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

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.

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*, pp. 3654–3663, 2021.

Zixuan Su, Xindi Shang, Jingjing Chen, Yu-Gang Jiang, Zhiyong Qiu, and Tat-Seng Chua. Video relation detection via multiple hypothesis association. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3127–3135, 2020.

Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10410–10419, 2021.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

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

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

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Sipeng Zheng, Shizhe Chen, and Qin Jin. Vrdformer: End-to-end video visual relation detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18836–18846, 2022.

A APPENDIX

A.1 OVERVIEW OF SUPPLEMENTARY MATERIAL

This supplementary material provides additional technical details and extended experimental analyses to support the proposed PACT framework. The contents are organized as follows:

- **Algorithm of PACT:** Provides a complete walkthrough of the PACT forward pass, including frame-level visual token extraction, pair-aware contextual token construction, and spatio-temporal relation modeling.
- **Implementation Details:** Describes how semantic text embeddings and entity tubelets are obtained from frozen pre-trained models, together with the loss weights and the span threshold used during inference.
- **Model Architecture:** Details the architectural configuration of the spatio-temporal transformer and the relation prediction heads.
- **Analysis on the Injected Information:** Analyzes which information should be injected into the shared visual tokens by comparing semantic text embeddings with learnable category tokens, learnable role tokens, and role indicators.
- **Analysis on Per-Predicate Recall:** Extends the per-predicate comparison presented in the main paper to the complete predicate set under both Recall@50 and Recall@100.
- **Qualitative Results:** Presents additional visualizations illustrating how full-frame contextual tokens contribute to relation prediction.
- **Limitations and Future Works:** Discusses the current limitations of PACT and outlines directions for future research.

A.2 ALGORITHM OF PACT

Algorithm 1 summarizes the forward procedure of PACT for a single ordered subject–object pair (i, j). The frame-level visual tokens are extracted once per video by the frozen image encoder and shared across all entity pairs. For each pair, these shared tokens are specialized by injecting the spatial overlap scores and the category text embeddings of the target subject and object, producing pair-aware contextual tokens. The tokens are then resampled to a fixed length with positional encodings derived from the original frame indices, processed by the spatio-temporal transformer, and decoded into predicate labels and temporal spans through the relation queries. Since the injection is pair-specific, Steps 2 and 3 are repeated per pair, while Step 1 is computed only once.

A.3 IMPLEMENTATION DETAILS

A.3.1 ENTITY TUBELET EXTRACTION

Semantic Text Embeddings. The category text embeddings injected into the visual tokens are obtained from the text encoder of a frozen Grounding DINO (Liu et al., 2024), which is also used to obtain the entity tubelets described below. Each category is encoded independently as a single caption, so that the resulting embedding is not affected by the other categories in the vocabulary, and underscores in the category names are replaced by spaces so that they are read as natural language. The embedding of a category is obtained by mean-pooling the token features of its name, excluding special tokens, followed by ℓ_2 normalization. These embeddings are computed once before training and kept fixed, and are projected into the H-dimensional hidden space by a learnable linear layer.

Object Detection. We use the same frozen Grounding DINO (Liu et al., 2024) to obtain per-frame detections, without any fine-tuning on the target benchmarks. The object categories of each benchmark are concatenated into a text prompt, and detections are obtained with a box threshold of 0.1 and a text threshold of 0.2, keeping at most 50 detections per frame. Since some category names are rarely observed as natural language expressions, we additionally provide common synonyms in the prompt and map the detected phrases back to the original category. Duplicate detections arising from synonyms are merged by class-wise non-maximum suppression with an IoU threshold of 0.7.

Algorithm 1 Forward procedure of PACT

Require: Frames $\{I_t\}_{t=1}^{T}$; tubelets $\mathcal{T}_i, \mathcal{T}_j$ with categories c_i, c_j; resampling length L
Ensure: Predicate distribution and frame-level relation confidence for each relation query

Step 1: Frame-level Visual Tokens ▷ Computed once per video
for $t = 1, \dots, T$ **do**
  Extract visual tokens from I_t using the frozen image encoder
  Project and normalize them into $\mathbf{Z}_t \in \mathbb{R}^{M \times H}$
end for

Step 2: Token Construction ▷ Computed per subject-object pair (i, j)
$\tau_{ij} \leftarrow \tau_i \cap \tau_j$
Encode c_i, c_j via the text encoder to obtain semantic embeddings $\mathbf{v}^s, \mathbf{v}^o \in \mathbb{R}^H$
for $t \in \tau_{ij}$ **and** $m = 1, \dots, M$ **do**
  $\rho^s_{t,m},\ \rho^o_{t,m} \leftarrow$ spatial overlap of token region Ω_m with bounding boxes $b_{i,t},\ b_{j,t}$
  $\hat{\mathbf{z}}_{t,m} \leftarrow \mathrm{LN}\big(\mathbf{z}_{t,m} + \gamma_p \phi(\boldsymbol{\rho}_{t,m}) + \gamma_c(\rho^s_{t,m}\mathbf{v}^s + \rho^o_{t,m}\mathbf{v}^o)\big)$
end for

Step 3: Relation Modeling and Decoding
Resample the token sequence $\hat{\mathbf{Z}}_{ij}$ to fixed length L
Add temporal positional encodings derived from original frame indices to obtain $\widetilde{\mathbf{Z}}_{ij}$
for $\ell = 1, \dots, L$ **do** ▷ Spatial interaction within each frame
  $\bar{\mathbf{Z}}_{ij}[\ell, :] \leftarrow \mathrm{Attn_s}(\widetilde{\mathbf{Z}}_{ij}[\ell, :])$
end for
for $m = 1, \dots, M$ **do** ▷ Temporal dynamics across frames
  $\mathbf{F}_{ij}[:, m] \leftarrow \mathrm{Attn_t}(\bar{\mathbf{Z}}_{ij}[:, m])$
end for
Refine relation queries $\mathbf{Q}_{ij}$ through the relation decoder using $\mathbf{F}_{ij}$ to obtain $\hat{\mathbf{Q}}_{ij}$
Predict predicate distributions and frame-level confidences from each query in $\hat{\mathbf{Q}}_{ij}$

Tubelet Construction. Per-frame detections are linked into entity tubelets by the two-stage association procedure of ByteTrack (Zhang et al., 2022), where high- and low-confidence detections are associated in separate stages. Detections with a confidence above 0.4 are first matched to existing tracks with an IoU threshold of 0.3, and the remaining detections above 0.1 are then matched with a lower IoU threshold of 0.2 to recover occluded or low-confidence entities. Track positions are propagated by a linear motion model whose velocity is updated with a smoothing factor of 0.5, and a track is kept alive for up to 30 frames without a match. Gaps of at most 15 frames within a track are filled by linear interpolation of the bounding boxes, and tracks shorter than 3 frames are discarded. The object category of a track is determined by majority voting over its matched detections. Finally, the 16 tracks with the highest accumulated confidence are retained per video and used as the entity tubelets. All variants in our ablation studies share the identical tubelets, so that the comparison reflects only the design of the relation representation.

A.3.2 Training and Inference

Loss Weights. The weights of the training objective described in the main paper are set as follows. The predicate classification term is weighted by 3.0, and the temporal localization term is weighted by $\lambda_{\text{span}} = 1.0$. For the auxiliary losses, we use $\lambda_{\text{cls}} = 0.7$ for the object classification term and $\lambda_{\text{box}} = 2.0$ for the bounding-box regression term. The same weights are used for both benchmarks.

Span Threshold. At inference, the temporal span of each relation is recovered by grouping consecutive time steps whose frame-level relation confidence exceeds a predefined threshold, as described in the main paper. We set this threshold to 0.1 on `ImageNet-VidVRD` and 0.5 on `VidOR`, selected on the validation set of each benchmark.

A.4 Model Architecture

Model Configuration. The hidden dimension is set to $H = 256$. The pair-aware contextual tokens are resampled to a fixed length of $L = 32$ by linear interpolation over the co-occurrence interval, and

Table 4: **Ablation study on semantic text embedding.** All variants inject a vector into the visual tokens of the subject and object regions, and differ in the information it encodes.

Method	Relation Detection		Relation Tagging	
	mAP	R@50	P@1	P@5
Category token	29.36	17.39	73.00	55.08
Role token	30.77	18.80	74.50	56.75
Role indicator	31.12	19.21	73.50	56.15
Ours	**32.11**	**20.14**	**77.50**	**57.40**

the bounding boxes of the subject and object are interpolated in the same manner. The positional encodings added after resampling are sinusoidal functions of the original frame indices. The spatial and temporal encoders each contain two transformer blocks with four attention heads, followed by a three-layer decoder with $N = 10$ queries.

Relation Heads. The refined query representations are fed into three heads. The predicate head is a linear classifier over $|\mathcal{P}| + 1$ categories, including the no-relation category, and is applied at every decoder layer so that the outputs of the intermediate layers serve as auxiliary supervision. The subject and object heads are two-layer feed-forward networks followed by linear layers that predict the bounding box and the object category of each entity, which provide the auxiliary supervision described in the main paper.

A.5 Analysis on the Injected Information

The semantic text embedding injection specializes the shared full-frame visual tokens for a target subject–object pair. To examine what information the injected vector should encode, we compare four variants in Tab. 4, all of which share the identical architecture and inject a vector according to the spatial overlap with the subject and object regions. Category token replaces the text embeddings with learnable embeddings defined for each object category, which are randomly initialized and learned from scratch on the training set. Role token instead defines two learnable embeddings representing the subject and object roles, indicating which entity plays which role without specifying its category. Role indicator encodes the same role information as a binary vector, where a visual token is assigned $(1, 0)$ when it overlaps only the subject, $(0, 1)$ when it overlaps only the object, $(1, 1)$ when it overlaps both, and $(0, 0)$ otherwise, which is then projected into the hidden space by a linear layer. Ours uses the category text embeddings from the pre-trained text encoder.

Role token and Role indicator achieve comparable performance, indicating that simply marking which regions correspond to the subject and the object already provides a useful signal for relation prediction. Interestingly, Category token performs worst among all variants, even though it encodes strictly more information than the role-based ones. Since these embeddings are randomly initialized, the model has to learn the meaning of each category from the training data alone, and the injected vector may act as an arbitrary identifier rather than a meaningful description of the entity. In contrast, our semantic text embeddings are drawn from a pre-trained language space in which semantically related categories are already close to each other, so that the injected vector conveys not only the identity of the target entities but also their semantic relationships. Our method achieves the best performance across all metrics, demonstrating that the benefit comes from the semantic structure of the pre-trained text embeddings rather than from category information alone.

B Analysis on Per-Predicate Recall

The main paper compares per-predicate Recall@100 for the ten most frequent predicates. Here we extend this comparison to the entire predicate set of `ImageNet-VidVRD` in Fig. 7, where the predicates are sorted by their frequency in the training set. The improvement observed for the most frequent predicates is consistently reproduced under Recall@50 in Fig. 7(a), indicating that the proposed representation ranks correct relations higher rather than merely recovering them within a larger prediction budget. Many of the predicates on which our method improves, such as "*stand_behind*", "*walk_front*", and "*move_beneath*", describe the relative spatial arrangement or

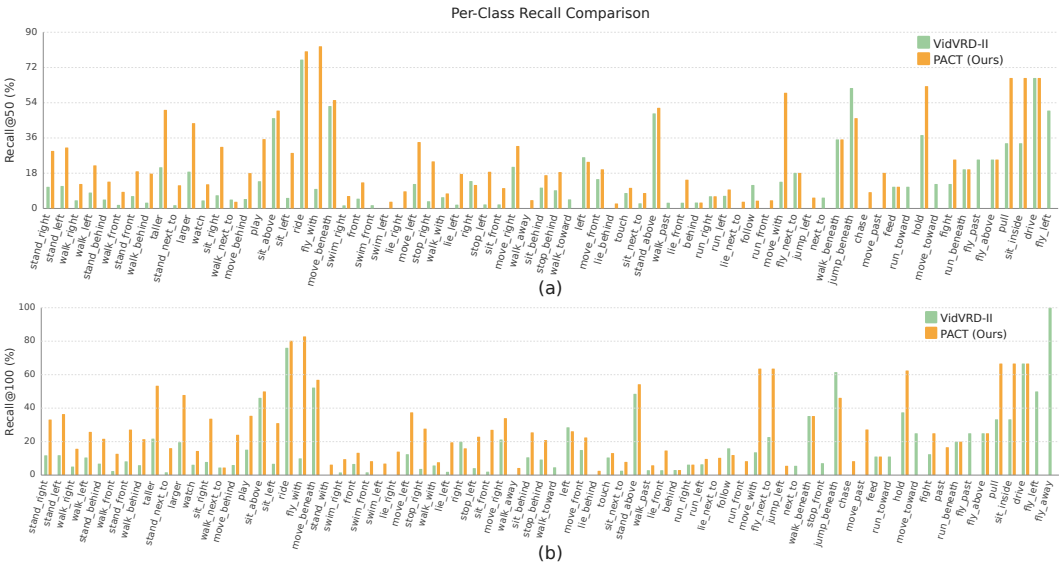

Figure 7: **Per-predicate recall for the entire predicate set on `ImageNet-VidVRD`.** (a) Recall@50 and (b) Recall@100. PACT (Ours) outperforms VidVRD-II (Shang et al., 2021) on most predicates under both metrics.

motion of the two objects, which is difficult to determine from the object regions alone. For the rare predicates in the tail of the distribution, both methods obtain low recall in Fig. 7(b), since few training instances are available for these categories. This suggests that the benefit of preserving full-frame context is largely orthogonal to the long-tailed predicate distribution, which remains a common challenge in VidVRD.

B.1 Qualitative Results

Additional qualitative results on `ImageNet-VidVRD` are shown in Fig. 8. Following the main paper, the top-5 relation predictions are compared with those of VidVRD-II (Shang et al., 2021), where each object is represented by the same color in its bounding box and the corresponding relation triplets. Check marks and crosses indicate correct and incorrect predictions, respectively. The sampled frames illustrate how the subject and object move within the scene over time, and the predicates that depend on this relative arrangement are predicted more reliably by our method.

B.2 Limitations and Future Works

B.2.1 Limitations

PACT constructs relation representations from entity tubelets produced by a pre-trained detector and tracker, and its performance is therefore bounded by the quality of these tubelets. Missed detections and identity switches propagate directly to the relation predictions, since a relation cannot be recovered once one of its entities is lost. Processing full-frame tokens for every candidate pair also increases the computational cost as the number of entities grows, which is more pronounced in crowded scenes.

Beyond the method itself, the evaluation protocol of current VidVRD benchmarks limits how accurately the improvement can be measured. A large portion of the predicate vocabulary is formed by combining an action with a relative direction, so that the correct label depends on the viewpoint from which the scene is observed rather than on the interaction itself. Moreover, the annotations cover only a subset of the relations that actually hold in a video, and a plausible prediction that is not annotated is counted as a false positive. As shown in our qualitative results, several predictions marked as incorrect are in fact reasonable descriptions of the scene, which suggests that the reported scores underestimate the actual quality of the predictions.



Figure 8: **Additional qualitative comparison of the top-5 relation predictions from VidVRD-II (Shang et al., 2021) and PACT on `ImageNet-VidVRD`.** Each object is represented by the same color in its bounding box and the corresponding relation triplets. Check marks and crosses indicate correct and incorrect predictions, respectively.

B.2.2 Future Works

Since PACT injects category semantics through a pre-trained text encoder rather than a fixed classifier, the injected vector is not restricted to a predefined category set. Together with Grounding DINO, which detects entities from arbitrary text queries, this design allows the entire pipeline to operate on categories that are described in natural language rather than enumerated in advance. Extending PACT toward open-vocabulary VidVRD is thus a natural next step, where both the entities and the predicates are represented in a shared language space so that unseen categories can be handled without retraining. We also plan to improve robustness through joint entity and relation modeling, and to reduce the computational cost by selecting the visual tokens that are most relevant to each pair.