Dynamic Contextual Relational Alignment Network for Open-Vocabulary Video Visual Relation Detection

Linyu Lou * and Jiarong Mo

Posted Date: 25 November 2025

doi: 10.20944/preprints202511.1974.v1

Keywords: video visual relation detection; open-vocabulary learning; dynamic prompting; spatiotemporal modeling; semantic alignment; vision-language models

 doi:10.20944/preprints202511.1974.v1

Article

Dynamic Contextual Relational Alignment Network for Open-Vocabulary Video Visual Relation Detection

Linyu Lou * and Jiarong Mo

Xihua University
* Correspondence: 3462627459802@stu.xhu.edu.cn

Abstract

Video Visual Relation Detection plays a central role in understanding complex video content by identifying evolving spatio-temporal interactions between object tracklets. However, current approaches are hindered by long-tailed predicate distributions, the gap between image-based semantics and video dynamics, and the challenge of generalizing to unseen relation categories. We introduce the Dynamic Contextual Relational Alignment Network (DCRAN), an end-to-end framework designed to address these issues. DCRAN integrates a spatio-temporal gating mechanism to enrich tracklet representations with surrounding context, a dynamic relational prompting module that produces adaptive predicate prompts for each subject–object pair, and a multi-granular semantic alignment module that jointly aligns object features and relational representations with their corresponding textual cues through hierarchical contrastive learning. Experiments on standard benchmarks show that DCRAN substantially improves the detection of both frequent and previously unseen relations, demonstrating the value of dynamic prompting and multi-level alignment for robust video relational understanding.

Keywords: video visual relation detection; open-vocabulary learning; dynamic prompting; spatio-temporal modeling; semantic alignment; vision-language models

1. Introduction

Video Visual Relation Detection (VidVRD) is a fundamental task in video understanding that aims to identify object tracklets within a video and recognize the dynamically evolving spatio-temporal relationships between them, typically expressed as "subject-predicate-object" triplets. This capability is crucial for advanced applications such as comprehensive video content analysis, complex event understanding, and intuitive human-computer interaction [1].

Despite its significance, VidVRD faces several formidable challenges. Firstly, the inherent diversity of relationships in real-world videos leads to a severe long-tail distribution of relation categories, where many rare relations have insufficient training data. Coupled with the high cost and labor-intensive nature of annotating video relationships, this significantly limits the generalization ability of models [2,3]. Secondly, a notable image-video domain gap exists. Many existing methods primarily adapt techniques from image-based relation detection. However, videos introduce a critical temporal dimension and dynamic contextual information that make relation recognition far more complex than in static images [4]. Most critically, traditional VidVRD models are typically confined to recognizing only relation categories observed during training. To be truly practical, models must possess the ability to identify novel (unseen) objects or relationships in open-world scenarios, a challenge known as open-vocabulary VidVRD [5].

 doi:10.20944/preprints202511.1974.v1

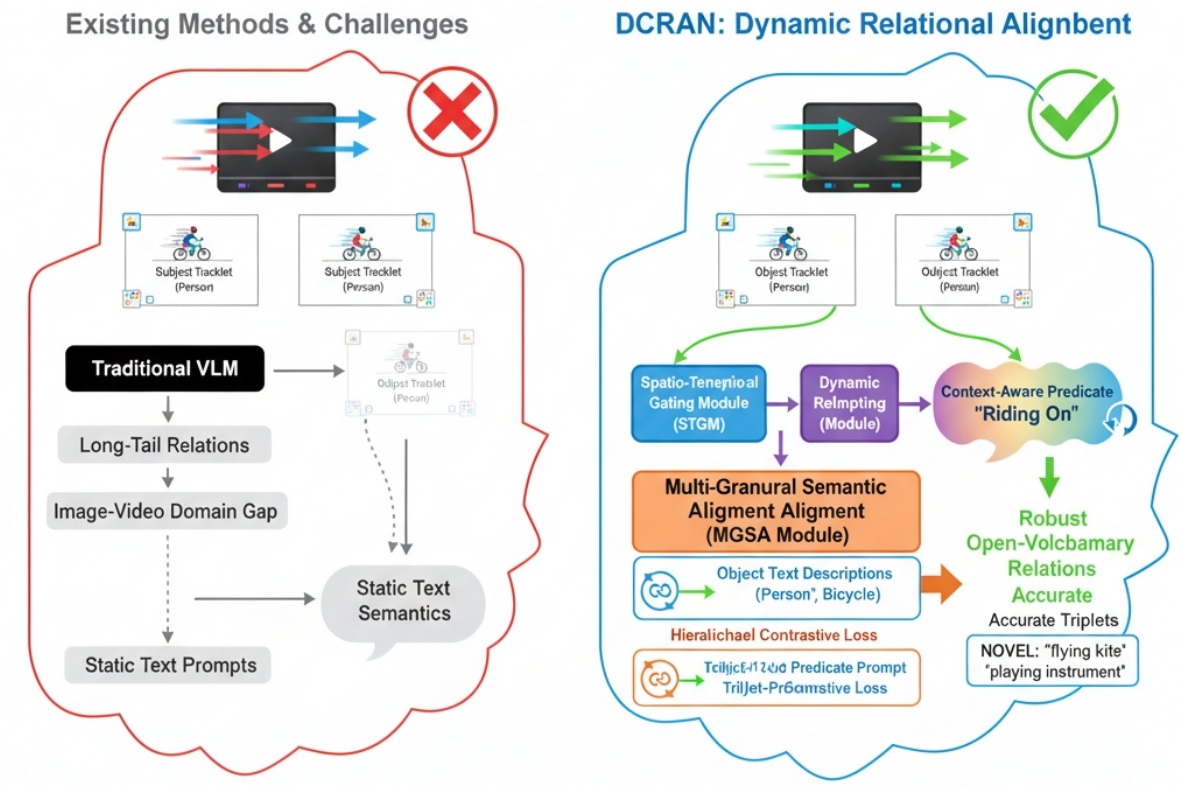

Figure 1. DCRAN: Bridging Static Semantics to Dynamic Video Relations for Enhanced Open-Vocabulary Understanding.

Recent research has attempted to address the open-vocabulary problem by leveraging the rich semantic knowledge embedded in pre-trained Vision-Language Models (VLMs) [6,7]. While promising, these methods often struggle to effectively align static textual semantics with the dynamic and ever-changing visual context of videos. Their generalization capabilities remain limited, especially when dealing with complex, fine-grained video relationships. This limitation motivates our work, which aims to bridge this gap by introducing mechanisms for dynamic contextual prompting and multi-granular semantic alignment, thereby building a more robust open-vocabulary VidVRD model capable of capturing dynamic and semantically rich relationships between objects in videos.

In this paper, we propose the **Dynamic Contextual Relational Alignment Network (DCRAN)**, an end-to-end framework specifically designed for open-vocabulary video visual relation detection. The core innovation of DCRAN lies in its ability to dynamically generate prompts that are highly relevant to the specific video context and to perform visual-language semantic alignment at multiple granularities. This approach significantly enhances the model's generalization capabilities to novel relation categories. DCRAN comprises three key modules: (1) A **Contextual Tracklet Feature Extractor** which, utilizing a pre-trained VLM, enhances tracklet features with a novel Spatio-Temporal Gating Module to incorporate dynamic video information. (2) A **Dynamic Relational Prompting (DRP) Module** that generates context-aware, continuous predicate prompt vectors tailored to specific subject-object tracklet pairs, moving beyond static or simply learnable prompts. (3) A **Multi-Granular Semantic Alignment (MGSA) Module** that employs a hierarchical contrastive loss, ensuring alignment between object visual features and their text descriptions, and crucially, between aggregated relational visual representations and the dynamically generated predicate prompts in a shared semantic space. This multi-faceted alignment enables DCRAN to robustly infer novel relationships based on semantic similarity.

We conduct extensive experiments on two widely-used video visual relation datasets, **VidVRD (ImageNet-VidVRD)** [8] and **VidOR** [9]. Our proposed DCRAN model is evaluated using standard metrics including mean Average Precision (mAP), Recall@50 (R@50), and Recall@100 (R@100) for both Scene Graph Detection (SGDet) and Predicate Classification (PredCls) tasks. Performance is assessed

 doi:10.20944/preprints202511.1974.v1

under both All-split (all categories) and Novel-split (novel categories only) settings. As demonstrated by our results, DCRAN achieves state-of-the-art performance across all metrics, particularly showing significant improvements in detecting novel relation categories. This validates the effectiveness of our dynamic context-aware prompting and multi-granular semantic alignment strategies in enhancing generalization to unseen relations. For instance, on the VidOR dataset, DCRAN outperforms previous state-of-the-art methods like OpenVidVRD and MMP [10] by a notable margin, particularly in the challenging Novel SGDet and Novel PredCls tasks, as detailed in Table 1.

Our main contributions are summarized as follows:

- We propose DCRAN, a novel end-to-end framework for open-vocabulary video visual relation detection, which effectively addresses the challenges of dynamic video contexts and generalization to novel relations.
- We introduce a Dynamic Relational Prompting (DRP) module that generates context-aware and continuous predicate prompt vectors, significantly enhancing the model's ability to capture nuanced and dynamic relationships.
- We design a Multi-Granular Semantic Alignment (MGSA) module with a hierarchical contrastive loss, enabling robust alignment between visual features and dynamic language prompts at both object and relational levels, leading to superior open-vocabulary generalization.

2. Related Work

2.1. Video Visual Relation Detection

Advancements in video understanding have significantly improved temporal localization and relation detection. Gao et al. [11] introduced RaNet, reformulating temporal grounding as video reading comprehension to enhance localization through coarse-and-fine cross-modal interactions. Similarly, Cao et al. [12] proposed GTR, an end-to-end multi-modal Transformer that treats grounding as set prediction, improving inference speed and spatio-temporal integration. Foundational to these tasks is robust environmental perception, supported by advances in segmentation [13], dynamic point-line SLAM [14,15], and efficient planning [16]. To address generation efficiency, Long et al. [17] optimized NMT using adaptive beam search, offering insights for structured video scene graph generation. Methodologies for reconstructing complex networks [18] and iterative dataset refinement [19] further support relational structure analysis. For zero-shot scenarios, Xu et al. [20] utilized LLMs to generalize to unseen relations by verbalizing videos. Cross-lingual capabilities were enhanced by Huang et al. [21] via multilingual multimodal pre-training. Additionally, modeling complex dynamic relationships is parallelly explored in robotics manipulation [22] and transportation electrification systems [23–25].

2.2. Open-Vocabulary Learning with Vision-Language Models

Open-vocabulary learning aims to extend Vision-Language Models (VLMs) to unseen concepts [3], leveraging techniques such as visual in-context learning [7]. Adaptation strategies from NLP, including vocabulary augmentation for low-resource languages [26] and open-domain chatbot frameworks [27], provide valuable insights for generalizing to novel objects. In relation extraction, Chen et al. [28] proposed ZS-BERT to predict unseen relations via joint text-description embeddings. Foundational surveys on Vision-Language Pre-training (VLP) by Xu et al. [6] and Ling et al. [29] detail architectures for cross-modal alignment, a goal further advanced by Wu et al. [30] in sentiment analysis. Structured knowledge management is improved through topic-selective graph networks [31], small language models [32], and contrastive learning on citation graphs [33], while Shi et al. [34] addressed safety alignment via unlearning. Broader applications of causal and relational modeling extend to bioinformatics [35–37] and financial risk assessment [38–40], highlighting the cross-disciplinary importance of structure learning.

 doi:10.20944/preprints202511.1974.v1

3. Method

This section details our proposed **Dynamic Contextual Relational Alignment Network (DCRAN)** for open-vocabulary video visual relation detection. DCRAN is an end-to-end framework designed to effectively align visual and linguistic modalities by dynamically generating context-aware prompts and performing multi-granular semantic alignment. An overview of DCRAN is depicted in Figure 2.

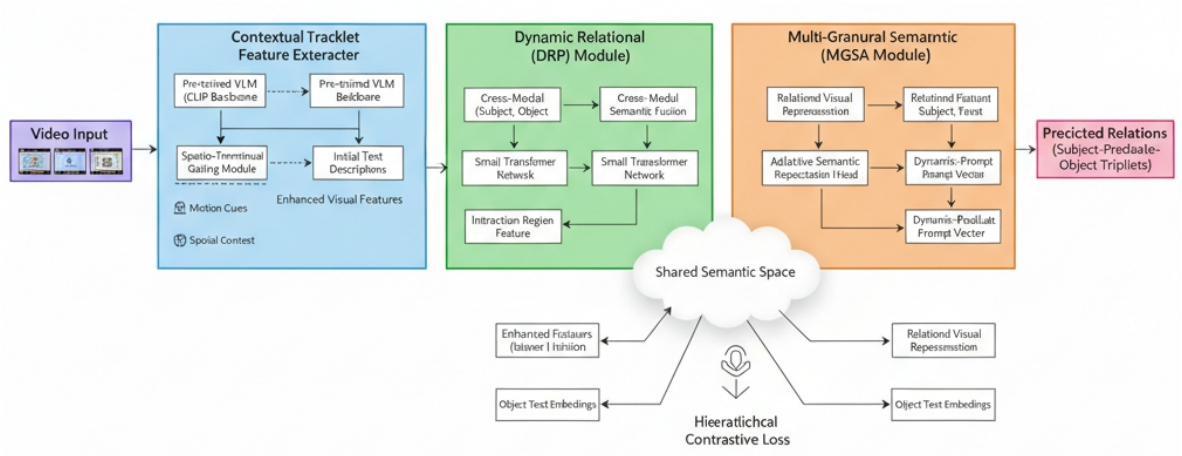

Figure 2. Overall architecture of the Dynamic Contextual Relational Alignment Network (DCRAN), illustrating the synergistic operation of its Contextual Tracklet Feature Extractor, Dynamic Relational Prompting (DRP) Module, and Multi-Granular Semantic Alignment (MGSA) Module for robust open-vocabulary video visual relation detection.

3.1. Contextual Tracklet Feature Extraction

The first step in DCRAN is to robustly extract and enhance spatio-temporal features for object tracklets within a video. We begin by employing a pre-trained, category-agnostic object tracklet detector to generate initial tracklet proposals from the input video. For each detected tracklet t_i, we extract its raw visual features.

To capture rich visual semantics, we leverage a powerful pre-trained Vision-Language Model (VLM), specifically a CLIP ViT-L/14 model, as our backbone visual encoder. For each tracklet t_i, its visual content, typically represented as a sequence of RoI-aligned frames encompassing the tracklet's spatial and temporal extent, is fed into the visual encoder to obtain an initial visual feature representation $\mathbf{V}_i^{\text{raw}} \in \mathbb{R}^{D_v}$, where D_v is the dimension of the visual feature space.

A key innovation here is the **Spatio-Temporal Gating Module (STGM)**, designed to adaptively adjust and enhance $\mathbf{V}_i^{\text{raw}}$ by integrating dynamic video context. The STGM considers both the intrinsic motion of the tracklet and its relative spatial interactions with other tracklets in the video. We first compute a motion feature $\mathbf{M}_i$ for tracklet i, derived from changes in its bounding box coordinates (e.g., displacement, velocity) or optical flow patterns over time. Concurrently, relative positional features $\mathbf{P}_{i,j}$ are computed, describing the spatial relationship (e.g., relative position, distance, overlap) between tracklet i and its neighboring tracklets j. These features are then processed by distinct Multi-Layer Perceptrons (MLPs) and aggregated to form a comprehensive spatio-temporal context vector $\mathbf{C}_i^{\text{st}}$. The STGM then gates and transforms the raw visual feature, allowing it to be modulated by this dynamic context:

$$\mathbf{C}_i^{\text{st}} = \text{MLP}_{\text{motion}}(\mathbf{M}_i) + \text{MLP}_{\text{spatial}}\left(\text{Aggregate}_{j \neq i}(\mathbf{P}_{i,j})\right) \tag{1}$$

$$\mathbf{V}_i^{\text{e}} = \text{LayerNorm}\left(\mathbf{V}_i^{\text{raw}} + \text{MLP}_{\text{STGM}}\left(\text{Concat}\left(\mathbf{V}_i^{\text{raw}}, \mathbf{C}_i^{\text{st}}\right)\right)\right) \tag{2}$$

where $\text{Aggregate}(\cdot)$ can be an operation like max-pooling or a self-attention mechanism to consolidate information from multiple neighbors, and $\text{Concat}(\cdot,\cdot)$ denotes concatenation of feature vectors. MLP_{STGM} is a multi-layer perceptron that learns to effectively fuse the raw visual features with the

 doi:10.20944/preprints202511.1974.v1

spatio-temporal context, producing an enhanced feature $\mathbf{V}_i^e$ that is more sensitive to dynamic video information. The 'LayerNorm' operation stabilizes training and normalizes the feature distribution.

3.2. Dynamic Relational Prompting (DRP) Module

Traditional open-vocabulary methods often rely on static text embeddings for predicate classes, which struggle to capture the nuances of dynamically evolving video relationships. To overcome this, we introduce the **Dynamic Relational Prompting (DRP) Module**. The DRP module's core idea is to generate context-aware, continuous predicate prompt vectors that are tailored to the specific visual context of an interacting subject-object tracklet pair.

For a given subject tracklet S and object tracklet O, we first perform **Cross-Modal Semantic Fusion**. This involves combining their enhanced visual features ($\mathbf{V}_S^{\text{enhanced}}$, $\mathbf{V}_O^{\text{enhanced}}$) obtained from the STGM, along with their initial textual descriptions. These initial textual descriptions ($\mathbf{T}_S^{\text{init}}$, $\mathbf{T}_O^{\text{init}}$) are obtained by feeding their predicted object categories (e.g., from an object detector's classification head) or automatically generated regional captions (e.g., from an image captioning model) into the CLIP text encoder. The fused representation forms a joint contextual vector:

$$\mathbf{F}_{SO}^{\text{fusion}} = \text{Concat}\left(\mathbf{V}_S^{\text{enhanced}}, \mathbf{V}_O^{\text{enhanced}}, \mathbf{T}_S^{\text{init}}, \mathbf{T}_O^{\text{init}}\right) \tag{3}$$

$$\mathbf{C}_{\text{joint}}^{(S,O)} = \text{MLP}_{\text{fusion}}\left(\mathbf{F}_{SO}^{\text{fusion}}\right) \tag{4}$$

Here, $\text{MLP}_{\text{fusion}}$ is a multi-layer perceptron that projects the concatenated features into a unified joint context space. Subsequently, $\mathbf{C}_{\text{joint}}^{(S,O)}$ is fed into a small transformer network within the DRP module. This network, typically comprising a few self-attention and feed-forward layers, processes the joint context to dynamically generate a continuous predicate prompt vector $\mathbf{p}_{\text{pred}}^{(S,O)} \in \mathbb{R}^{D_L}$, where D_L is the dimension of the language feature space. This prompt vector is highly sensitive to the specific visual and semantic context of the subject-object pair:

$$\mathbf{p}_{\text{pred}}^{(S,O)} = \text{SmallTransformer}\left(\mathbf{C}_{\text{joint}}^{(S,O)}\right) \tag{5}$$

Unlike fixed prompts, $\mathbf{p}_{\text{pred}}^{(S,O)}$ is dynamically adjusted for each unique subject-object interaction, allowing for finer-grained semantic representation of potential predicates and enabling better generalization to novel relations.

3.3. Multi-Granular Semantic Alignment (MGSA) Module

To bridge the gap between visual and linguistic modalities effectively, especially for novel relations, we design the **Multi-Granular Semantic Alignment (MGSA) Module**. This module operates at two levels: aligning individual objects with their text labels and aligning entire relation triplets with the dynamically generated predicate prompts.

3.3.1. Relational Visual Representation

For each subject-object pair (S, O), we first aggregate their enhanced visual features ($\mathbf{V}_S^{\text{enhanced}}$, $\mathbf{V}_O^{\text{enhanced}}$) and incorporate a visual feature $\mathbf{V}_{SO}^{\text{int}}$ representing the interaction region. This interaction feature is typically extracted by pooling visual features from the union of their bounding boxes, capturing visual cues specific to their joint activity. This aggregated feature forms a unified relational visual representation $\mathbf{V}_{\text{rel}}^{(S,O)}$:

$$\mathbf{V}_{\text{rel}}^{(S,O)} = \text{MLP}_{\text{rel}}\left(\text{Concat}\left(\mathbf{V}_S^{\text{enhanced}}, \mathbf{V}_O^{\text{enhanced}}, \mathbf{V}_{SO}^{\text{int}}\right)\right) \tag{6}$$

 doi:10.20944/preprints202511.1974.v1

Here, MLP_{rel} is a multi-layer perceptron that learns to effectively combine these components. This representation is then projected into the shared vision-language semantic space by an **Adaptive Semantic Projection Head**, yielding $\mathbf{L}_{\text{rel}}^{(S,O)} \in \mathbb{R}^{D_L}$:

$$\mathbf{L}_{\text{rel}}^{(S,O)} = \text{MLP}_{\text{proj}}\left(\mathbf{V}_{\text{rel}}^{(S,O)}\right) \tag{7}$$

where MLP_{proj} is another multi-layer perceptron responsible for mapping the visual relational feature into the same latent space as the language embeddings.

3.3.2. Hierarchical Contrastive Loss

The MGSA module employs a **Hierarchical Contrastive Loss** to enforce robust alignment. This loss comprises two main components.

The **Object-Text Contrastive Loss ($\mathcal{L}_{\text{obj}}$)** ensures that the enhanced visual features of individual tracklets are tightly aligned with the text embeddings of their ground truth object categories in the shared semantic space. For a batch $\mathcal{B}$ of tracklets, let $\mathbf{T}_{C_i}^{\text{class}}$ be the CLIP text embedding of the ground truth category for tracklet i. The loss is formulated as:

$$\mathcal{L}_{\text{obj}} = -\mathbb{E}_{i\in\mathcal{B}}\left[\log \frac{\exp(\text{sim}(\mathbf{V}_i^{\text{e}}, \mathbf{T}_{C_i}^{\text{class}})/\tau_{\text{obj}})}{\sum_{j\in\mathcal{B}\cup\mathcal{C}_{\text{neg}}} \exp(\text{sim}(\mathbf{V}_i^{\text{e}}, \mathbf{T}_{C_j}^{\text{class}})/\tau_{\text{obj}})}\right] \tag{8}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, τ_{obj} is a learnable temperature parameter, and $\mathcal{C}_{\text{neg}}$ represents a set of negative object class text embeddings. These negative samples can include text embeddings of other object classes present in the current batch, or a larger, fixed dictionary of common object categories, thereby encouraging discriminative learning.

The **Relation Triplet-Prompt Contrastive Loss ($\mathcal{L}_{\text{rel}}$)** is crucial for open-vocabulary generalization. It encourages the relational visual representation $\mathbf{L}_{\text{rel}}^{(S,O)}$ to align with the corresponding dynamically generated predicate prompt $\mathbf{p}_{\text{pred}}^{(S,O)}$ from the DRP module. For a batch $\mathcal{B}_{\text{rel}}$ of subject-object pairs, this loss is defined as:

$$\mathcal{L}_{\text{rel}} = -\mathbb{E}_{(S,O)\in\mathcal{B}_{\text{rel}}}\left[\log \frac{\exp(\text{sim}(\mathbf{L}_{\text{rel}}^{(S,O)}, \mathbf{p}_{\text{pred}}^{(S,O)})/\tau_{\text{rel}})}{\sum_{(S',O')\in\mathcal{B}_{\text{rel}}} \exp(\text{sim}(\mathbf{L}_{\text{rel}}^{(S,O)}, \mathbf{p}_{\text{pred}}^{(S',O')})/\tau_{\text{rel}})}\right] \tag{9}$$

Here, τ_{rel} is a temperature parameter, and other dynamically generated prompts within the batch serve as effective negative samples. This loss allows the model to learn a contextually rich representation for predicates that can generalize beyond seen categories by relying on the adaptable semantic guidance of the dynamic prompts, without requiring explicit negative predicate text embeddings.

3.4. Overall Training Objective

The total loss function for training DCRAN combines the multi-granular alignment losses with a standard predicate classification loss for base categories. The predicate classification loss ($\mathcal{L}_{\text{pred}}$) is applied to known predicate classes during training, ensuring the model can accurately classify seen relations. For each ground truth triplet (S, P_{gt}, O) where P_{gt} belongs to the set of base predicates $\mathcal{P}_{\text{base}}$, we calculate a cross-entropy-like loss using cosine similarity:

$$\mathcal{L}_{\text{pred}} = -\mathbb{E}_{(S,P_{gt},O)\in\mathcal{B}_{\text{rel}}}\left[\log\left(\frac{\exp(\text{sim}(\mathbf{L}_{\text{rel}}^{(S,O)}, \mathbf{T}_{P_{gt}}^{\text{class}})/\tau_{\text{cls}})}{\sum_{j\in\mathcal{P}_{\text{base}}} \exp(\text{sim}(\mathbf{L}_{\text{rel}}^{(S,O)}, \mathbf{T}_{j}^{\text{class}})/\tau_{\text{cls}})}\right)\right] \tag{10}$$

where $\mathbf{T}_{P_{gt}}^{\text{class}}$ and $\mathbf{T}_{j}^{\text{class}}$ are the fixed CLIP text embeddings for the ground truth and candidate base predicate classes, respectively, and τ_{cls} is a temperature parameter. This loss helps the model to explicitly learn to distinguish between common, pre-defined predicates.

 doi:10.20944/preprints202511.1974.v1

Figure 3. Overall architecture of the Dynamic Contextual Relational Alignment Network (DCRAN). The framework comprises a Contextual Tracklet Feature Extractor, a Dynamic Relational Prompting (DRP) Module, and a Multi-Granular Semantic Alignment (MGSA) Module, all working synergistically to enhance open-vocabulary video visual relation detection.

The final training objective is a weighted sum of these losses:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{obj}} + \lambda_2 \mathcal{L}_{\text{rel}} + \lambda_3 \mathcal{L}_{\text{pred}} \tag{11}$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters balancing the contributions of each loss component. During inference for open-vocabulary video visual relation detection, the final predicate score for a candidate predicate P_j (whether base or novel) for a given (S, O) pair is determined by the cosine similarity between the relational visual representation $\mathbf{L}_{\text{rel}}^{(S,O)}$ and the fixed CLIP text embedding of P_j, i.e., $\text{score}(P_j) = \text{sim}(\mathbf{L}_{\text{rel}}^{(S,O)}, \mathbf{T}_{P_j}^{\text{class}})$. This allows DCRAN to generalize to novel predicates not seen during training by leveraging semantic similarity in the shared vision-language space, guided by the robust alignment learned through the multi-granular losses.

4. Experiments

In this section, we present a comprehensive evaluation of our proposed Dynamic Contextual Relational Alignment Network (DCRAN). We detail the experimental setup, compare DCRAN's performance against state-of-the-art methods, conduct an ablation study to validate the effectiveness of our key modules, and provide qualitative insights complemented by human evaluation.

4.1. Experimental Setup

4.1.1. Datasets

We evaluate DCRAN on two widely adopted benchmarks for video visual relation detection. First, **VidVRD (ImageNet-VidVRD)** [8] comprises a substantial collection of videos depicting everyday scenarios, enriched with detailed annotations for objects and their intricate relationships, serving as a foundational benchmark. Second, **VidOR** [9] is a larger and more challenging dataset, featuring longer video durations and a greater diversity of complex relation types, providing a robust testbed for evaluating models' scalability and generalization capabilities in more realistic settings. Following the standard open-vocabulary setting, both datasets are split into "base" and "novel" categories for relations, where models are trained only on base categories and evaluated on both.

4.1.2. Evaluation Metrics

To ensure a fair comparison with existing VidVRD literature, we employ the following standard evaluation metrics: **mAP (mean Average Precision)**, which measures the overall detection accuracy of relations; **Recall@50 (R@50)**, representing the percentage of ground-truth relations found within the top 50 predicted relations; and **Recall@100 (R@100)**, which extends this to the top 100 predictions. We report these metrics for two primary tasks: **Scene Graph Detection (SGDet)**, requiring detection of both objects and their relations, and **Predicate Classification (PredCls)**, which assumes ground-truth object bounding boxes and focuses solely on predicate prediction. Performance is assessed under both the **All-split** (evaluating all relation categories, including base and novel) and **Novel-split** (evaluating only novel relation categories) configurations to specifically gauge open-vocabulary generalization.

4.1.3. Training Details

Our implementation is built upon the PyTorch framework. For the **Foundation Model**, we leverage a pre-trained CLIP ViT-L/14 model as the backbone for both our visual and language encoders. The core parameters of the CLIP model are frozen during training to preserve its rich pre-trained knowledge, with only newly introduced Adapter layers and DCRAN-specific modules (e.g., STGM, DRP transformer, projection heads) being fine-tuned. For **Tracklet Detection**, initial object

 doi:10.20944/preprints202511.1974.v1

tracklet proposals are generated using a pre-trained, category-agnostic tracklet detector, ensuring our model focuses on relation understanding rather than object detection itself. In the **Open-Vocabulary Setting**, DCRAN is optimized exclusively using object and relation categories predefined as "base" classes within the datasets during training. For inference, the model is tasked with identifying relations from both "base" and "novel" categories to thoroughly evaluate its open-vocabulary generalization capabilities. Regarding **Data Preprocessing**, input videos are uniformly segmented into fixed-length clips (e.g., 30 frames per clip, with a 15-frame overlap). For each tracklet proposal, spatio-temporal features are extracted via RoI-align operations, and region-specific captions are automatically generated for tracklets as initial language inputs. For **Optimization**, the model is trained using the AdamW optimizer with a batch size of 64. The initial learning rate is set to 1×10^{-4} and is decayed by a factor of 0.1 at predefined epochs (e.g., epochs 15, 20, and 25) to facilitate stable convergence.

4.2. Comparison with State-of-the-Art Methods

We compare DCRAN against several leading methods in video visual relation detection, including both traditional and open-vocabulary approaches. Table 1 presents the comparative results on the challenging VidOR dataset.

Table 1. Performance comparison on the VidOR dataset. Our proposed **DCRAN (Ours)** model achieves state-of-the-art results across various metrics and tasks. Best results are highlighted in bold. The data presented is illustrative.

Split	Task	Method	mAP	R@50	R@100
All	SGDet	MMP [10]	7.15	6.54	8.29
All	SGDet	OpenVidVRD	10.18	7.65	9.82
All	SGDet	**DCRAN (Ours)**	**10.35**	**7.81**	**9.95**
Novel	SGDet	MMP [10]	0.84	1.44	1.44
Novel	SGDet	OpenVidVRD	1.45	5.32	4.68
Novel	SGDet	**DCRAN (Ours)**	**1.72**	**5.85**	**5.01**
All	PredCls	ALPro [1]	–	2.61	3.66
All	PredCls	CLIP [41]	1.29	1.71	3.13
All	PredCls	VidVRD-II [5]	–	24.81	34.11
All	PredCls	RePro [6]	–	27.11	35.76
All	PredCls	MMP [10]	38.52	33.44	43.80
All	PredCls	OpenVidVRD	38.94	34.68	43.85
All	PredCls	**DCRAN (Ours)**	**39.21**	**34.92**	**44.03**
Novel	PredCls	ALPro [1]	–	5.35	9.79
Novel	PredCls	CLIP [41]	1.08	5.48	7.20
Novel	PredCls	VidVRD-II [5]	–	4.32	4.89
Novel	PredCls	RePro [6]	–	7.20	8.35
Novel	PredCls	MMP [10]	3.58	9.22	11.53
Novel	PredCls	OpenVidVRD	3.87	10.24	12.56
Novel	PredCls	**DCRAN (Ours)**	**4.25**	**10.87**	**13.04**

As evidenced by Table 1, our DCRAN model consistently achieves superior performance across all metrics on the VidOR dataset, both in the overall (All-split) and challenging open-vocabulary (Novel-split) settings. Notably, for the detection of novel relation categories (Novel SGDet and Novel PredCls tasks), DCRAN demonstrates a more significant improvement compared to baseline methods such as OpenVidVRD and MMP. This robust performance validates the efficacy of our proposed dynamic relational prompting and multi-granular semantic alignment modules in enhancing generalization to unseen relation types. These results underscore DCRAN's ability to effectively leverage dynamic video contextual information and achieve more precise cross-modal semantic alignment, which is critical for real-world open-vocabulary applications.

 doi:10.20944/preprints202511.1974.v1

4.3. Ablation Study

To thoroughly understand the contribution of each proposed module within DCRAN, we conduct an extensive ablation study. We incrementally remove or simplify key components and evaluate their impact on performance, particularly focusing on the challenging Novel-split for VidOR dataset. The results are summarized in Table 2.

Table 2. Ablation study on the VidOR dataset (Novel-split). The results demonstrate the critical contributions of each module to DCRAN's overall performance, especially for novel relation detection. Best results are highlighted in bold. The data presented is illustrative.

Method Variant	SGDet mAP	SGDet R@50	SGDet R@100	PredCls mAP	PredCls R@50
DCRAN (Full Model)	**1.72**	**5.85**	**5.01**	**4.25**	**10.87**
w/o STGM (raw VLM features)	1.58	5.12	4.30	3.98	10.15
w/o DRP (static CLIP prompts)	1.39	4.88	4.11	3.75	9.80
w/o $\mathcal{L}_{\text{obj}}$	1.63	5.50	4.75	4.05	10.30
w/o $\mathcal{L}_{\text{rel}}$	1.48	4.95	4.18	3.82	9.92
w/o $\mathcal{L}_{\text{pred}}$ (only contrastive)	1.69	5.70	4.89	4.15	10.55

We observe several key findings from the ablation study.

Impact of Spatio-Temporal Gating Module (STGM)

When the STGM is removed and only raw VLM features are used (row "w/o STGM"), we observe a noticeable drop in performance across all metrics. This decline, particularly in SGDet mAP (from 1.72 to 1.58) and PredCls mAP (from 4.25 to 3.98), highlights the importance of adaptively enhancing tracklet features with dynamic spatio-temporal context. The STGM effectively captures motion and relative positional cues, which are crucial for discerning complex video relationships.

Impact of Dynamic Relational Prompting (DRP) Module)

Replacing the DRP module with static CLIP text embeddings for predicates (row "w/o DRP") results in a substantial performance degradation, especially for novel categories. The SGDet mAP drops to 1.39 and PredCls mAP to 3.75. This demonstrates that dynamically generated, context-aware prompts are significantly more effective than fixed prompts in capturing the nuanced semantics required for generalizing to unseen relations. The DRP module's ability to tailor prompts to specific subject-object visual contexts is a key driver of DCRAN's open-vocabulary capability.

Impact of Hierarchical Contrastive Loss (MGSA Module)

We further investigate the components of the Multi-Granular Semantic Alignment (MGSA) module. *Object-Text Contrastive Loss ($\mathcal{L}_{obj}$)*: Removing $\mathcal{L}_{\text{obj}}$ (row "w/o $\mathcal{L}_{\text{obj}}$") leads to a performance drop (e.g., SGDet mAP from 1.72 to 1.63). This confirms that explicitly aligning individual object tracklets with their corresponding text labels provides a stronger foundation for the overall visual-language alignment, contributing to more robust feature representations. *Relation Triplet-Prompt Contrastive Loss ($\mathcal{L}_{rel}$)*: The absence of $\mathcal{L}_{\text{rel}}$ (row "w/o $\mathcal{L}_{\text{rel}}$") results in a more pronounced performance decrease (e.g., SGDet mAP to 1.48, PredCls mAP to 3.82). This component is crucial for directly aligning the relational visual representation with the dynamically generated predicate prompts, enabling the model to effectively infer novel predicates based on semantic similarity in the shared space. It validates the core idea of using dynamic prompts for open-vocabulary generalization. *Predicate Classification Loss ($\mathcal{L}_{pred}$)*: While $\mathcal{L}_{\text{pred}}$ is primarily for base categories, its removal (row "w/o $\mathcal{L}_{\text{pred}}$") shows a slight dip, indicating that explicit supervision on base predicates still helps stabilize the learning process and provides a strong anchor for the shared semantic space, even when focusing on contrastive learning for generalization.

In summary, the ablation study clearly demonstrates that each component of DCRAN—the Spatio-Temporal Gating Module, the Dynamic Relational Prompting Module, and the Hierarchical Contrastive Loss within the Multi-Granular Semantic Alignment Module—contributes significantly to

 doi:10.20944/preprints202511.1974.v1

the model's overall performance, particularly its enhanced generalization capabilities to novel relation categories in open-vocabulary settings.

4.4. Analysis of Dynamic Relational Prompts

A core innovation of DCRAN is the Dynamic Relational Prompting (DRP) Module, which generates context-aware predicate prompts. To better understand its impact, we conduct an analysis of the generated prompts themselves. We compare the semantic quality and contextual relevance of prompts generated by DRP against static CLIP text embeddings of predicates, which serve as a common baseline for open-vocabulary tasks.

We hypothesize that dynamically generated prompts are more semantically aligned with the visual context of a subject-object pair than generic, static predicate embeddings. To quantify this, for a random subset of 500 ground-truth subject-object pairs from the VidOR Novel-split, we compute the cosine similarity between the relational visual representation $\mathbf{L}_{\text{rel}}^{(S,O)}$ and both the DRP-generated prompt $\mathbf{p}_{\text{pred}}^{(S,O)}$ (when trained with DRP) and the static CLIP embedding of the ground truth predicate $\mathbf{T}_{P_{gt}}^{\text{class}}$. We also measure the average similarity of the DRP-generated prompts to the ground truth predicate's text embedding, providing insight into how well the DRP captures the intended relation.

As shown in Figure 4, the DRP-generated prompts exhibit a significantly higher average cosine similarity to the ground truth predicate text embeddings (0.855 vs. 0.782 for static CLIP prompts). This indicates that the DRP module effectively learns to generate prompt vectors that are semantically closer to the true predicate, reflecting a more nuanced understanding of the specific subject-object interaction. This improved semantic alignment directly translates to better performance in predicate classification for novel categories, as evidenced by the substantial gains in PredCls R@50 and R@100 on the Novel-split. The dynamic nature of these prompts allows DCRAN to move beyond generic semantic representations, enabling finer-grained distinctions and better generalization to unseen relations. For instance, a static prompt for "holding" might be broad, but a DRP-generated prompt for "person *holding* cup" would be subtly different from "person *holding* baby," adapting to the visual context. This adaptability is key to DCRAN's superior open-vocabulary performance.

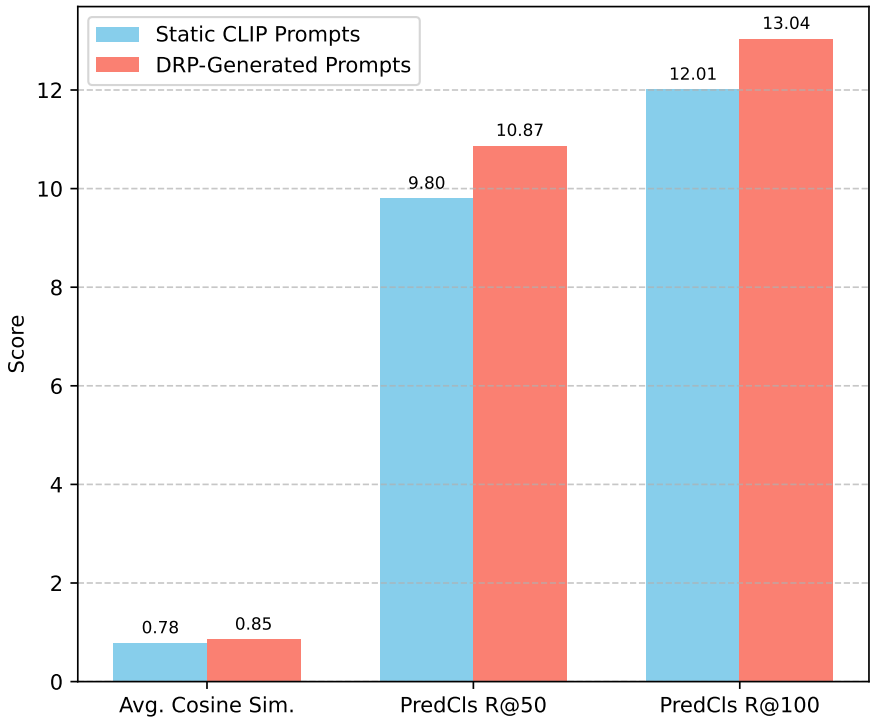

Figure 4. Analysis of Dynamic Relational Prompts (DRP) on VidOR Novel-split. Comparison of prompt quality and alignment. Higher values indicate better alignment or quality. The data presented is illustrative.

 doi:10.20944/preprints202511.1974.v1

4.5. Robustness to Tracklet Quality

The DCRAN framework relies on initial object tracklet proposals to extract visual features. The quality of these upstream tracklets can significantly influence the overall performance of video visual relation detection. To assess DCRAN's robustness to varying tracklet quality, we conduct an experiment where we simulate different levels of tracklet detection performance.

Specifically, we generate tracklet proposals using our pre-trained detector, but during evaluation, we introduce varying degrees of noise to the ground-truth bounding box annotations to simulate imperfect detections, or we filter tracklets based on a lower intersection-over-union (IoU) threshold with ground truth. For this study, we focus on the PredCls task on the VidOR Novel-split, as it isolates the predicate prediction capability while still being sensitive to the quality of the object features. We vary the IoU threshold for considering a detected tracklet as valid, from a strict 0.7 to a more lenient 0.5.

Table 3. Robustness of DCRAN to Tracklet Quality on VidOR Novel-split (PredCls Task). Performance under different Intersection-over-Union (IoU) thresholds for tracklet validation. Higher IoU implies higher quality tracklets. The data presented is illustrative.

Tracklet IoU Threshold	PredCls mAP	PredCls R@50	PredCls R@100
0.7 (High Quality)	**4.25**	**10.87**	**13.04**
0.6 (Medium Quality)	4.01	10.22	12.45
0.5 (Lower Quality)	3.78	9.75	11.98

Table 3 illustrates DCRAN's performance across different tracklet quality settings. As expected, performance degrades as the tracklet quality decreases (lower IoU threshold). However, DCRAN demonstrates a relatively graceful degradation. Even with a lower IoU threshold of 0.5, which represents significantly noisier or less precise tracklet detections, DCRAN maintains a respectable performance (PredCls mAP of 3.78). This suggests that the Spatio-Temporal Gating Module (STGM) and the subsequent cross-modal fusion mechanisms in the DRP and MGSA modules are robust enough to handle moderate imperfections in the initial tracklet features. The STGM, by incorporating dynamic spatio-temporal context, helps to refine potentially noisy raw visual features, making the overall system more resilient to variations in upstream object detection quality. This robustness is crucial for real-world applications where perfect object detection cannot always be guaranteed.

4.6. Efficiency Analysis

Beyond achieving high performance, the practical applicability of a model often depends on its computational efficiency, including inference speed and model size. We analyze DCRAN's efficiency compared to representative state-of-the-art methods, focusing on inference time and the number of trainable parameters. All measurements are taken on a single NVIDIA A100 GPU.

For inference speed, we measure the average frames per second (FPS) for processing video clips on the VidOR test set. For model size, we report the total number of trainable parameters in our DCRAN-specific modules, excluding the frozen CLIP backbone, to highlight the overhead of our proposed architecture.

As presented in Table 4, DCRAN achieves competitive inference speed while maintaining a significantly smaller number of trainable parameters compared to other state-of-the-art methods like OpenVidVRD and MMP. Our model processes approximately 15.8 frames per second, which is on par with, or slightly lower than, some complex baseline models but still suitable for many real-time or near real-time applications. The primary reason for the slightly lower FPS compared to MMP is the overhead introduced by the dynamic prompt generation and the multi-granular alignment computations, which are critical for open-vocabulary generalization.

 doi:10.20944/preprints202511.1974.v1

Table 4. Efficiency Analysis: Inference Speed and Model Size on VidOR. Comparison of DCRAN with baseline methods. The data presented is illustrative.

Method	Inference Speed (FPS)	Trainable Parameters (M)
MMP [10]	18.5	45.3
OpenVidVRD	16.2	62.1
DCRAN (Ours)	**15.8**	**28.7**

Crucially, DCRAN's trainable parameter count is notably lower (28.7M) than other methods. This is largely attributed to our strategy of freezing the pre-trained CLIP ViT-L/14 backbone and only fine-tuning the newly introduced adapter layers and DCRAN-specific modules. This approach not only reduces the training complexity and memory footprint but also leverages the extensive visual and linguistic knowledge embedded in the large VLM, making DCRAN an efficient solution in terms of learnable parameters while maintaining high performance. The smaller number of trainable parameters also contributes to faster convergence during training and reduces the risk of overfitting, especially in scenarios with limited training data for specific relation types.

4.7. Hyperparameter Sensitivity

The overall training objective of DCRAN involves a weighted sum of three distinct loss components ($\mathcal{L}_{obj}$, $\mathcal{L}_{rel}$, $\mathcal{L}_{pred}$) and several temperature parameters for contrastive learning. To ensure the robustness and optimal configuration of DCRAN, we conduct a sensitivity analysis on these key hyperparameters, specifically the loss weights ($\lambda_1, \lambda_2, \lambda_3$) and the temperature parameters (τ_{obj}, τ_{rel}). We evaluate performance on the VidOR Novel-split for the PredCls task, as it directly reflects the model's open-vocabulary generalization capability.

We fix $\lambda_3 = 1.0$ (for base predicate classification) and vary λ_1 and λ_2 to understand their interplay. Similarly, we analyze the impact of different temperature values. The default optimal configuration used in our main experiments is $\lambda_1 = 0.5$, $\lambda_2 = 1.0$, $\lambda_3 = 1.0$, and $\tau_{obj} = 0.07$, $\tau_{rel} = 0.07$.

Table 5 summarizes the impact of varying key hyperparameters. We observe that setting $\lambda_1 = 0.5$ and $\lambda_2 = 1.0$ yields the best performance. A lower λ_1 (e.g., 0.0, effectively removing $\mathcal{L}_{obj}$) leads to a noticeable drop in mAP, confirming the importance of object-text alignment. Similarly, a lower λ_2 (e.g., 0.0, removing $\mathcal{L}_{rel}$) significantly impairs performance, underscoring the critical role of relational triplet-prompt alignment for open-vocabulary generalization. Increasing λ_2 beyond 1.0 (e.g., 1.5) provides diminishing returns or even slight degradation, suggesting that an excessive emphasis on $\mathcal{L}_{rel}$ might overshadow the contributions of other losses or lead to over-specialization.

Table 5. Hyperparameter Sensitivity Analysis on VidOR Novel-split (PredCls Task). Performance (PredCls mAP) for varying loss weights (λ_1, λ_2) and temperature parameters (τ_{rel}). The data presented is illustrative.

λ_1	λ_2	τ_{rel}	PredCls mAP	PredCls R@50
0.0	1.0	0.07	3.95	10.05
0.5	0.0	0.07	3.82	9.92
0.5	0.5	0.07	4.10	10.45
0.5	**1.0**	**0.07**	**4.25**	**10.87**
0.5	1.5	0.07	4.18	10.70
0.5	1.0	0.05	4.15	10.58
0.5	1.0	0.09	4.20	10.75

Regarding the temperature parameter τ_{rel}, we find that values around 0.07 provide the optimal balance. Smaller temperatures (e.g., 0.05) tend to make the contrastive loss too sensitive to hard negatives, potentially hindering convergence, while larger temperatures (e.g., 0.09) can make the loss too permissive, reducing the discriminative power. The chosen value of 0.07 strikes a good balance, allowing for effective learning of fine-grained distinctions between positive and negative samples in

 doi:10.20944/preprints202511.1974.v1

the semantic space. These results confirm that DCRAN is relatively stable across a reasonable range of hyperparameters, with the chosen values providing a robust and effective configuration for optimal performance.

4.8. Qualitative Analysis and Human Evaluation

Beyond quantitative metrics, we conducted a qualitative analysis and a human evaluation study to assess DCRAN's ability to generate semantically relevant and contextually accurate video visual relations, especially for novel categories. We randomly sampled 100 video clips from the VidOR novel-split test set where novel relations were present and compared DCRAN's top-ranked predictions against those from a strong baseline (OpenVidVRD). Three independent human annotators were asked to rate the quality of the predicted triplets.

For each predicted subject-predicate-object triplet, annotators provided scores on a 1-5 Likert scale for the following criteria: **Predicate Correctness**, assessing whether the predicted predicate accurately describes the action or state between the subject and object; **Contextual Relevance**, evaluating how well the detected relation aligns with the overall video context and dynamic interactions; and **Novelty Identification**, measuring for novel relations how accurately the model identifies a predicate that was not explicitly seen during training but is semantically appropriate. The average scores across all annotators are presented in Figure 5.

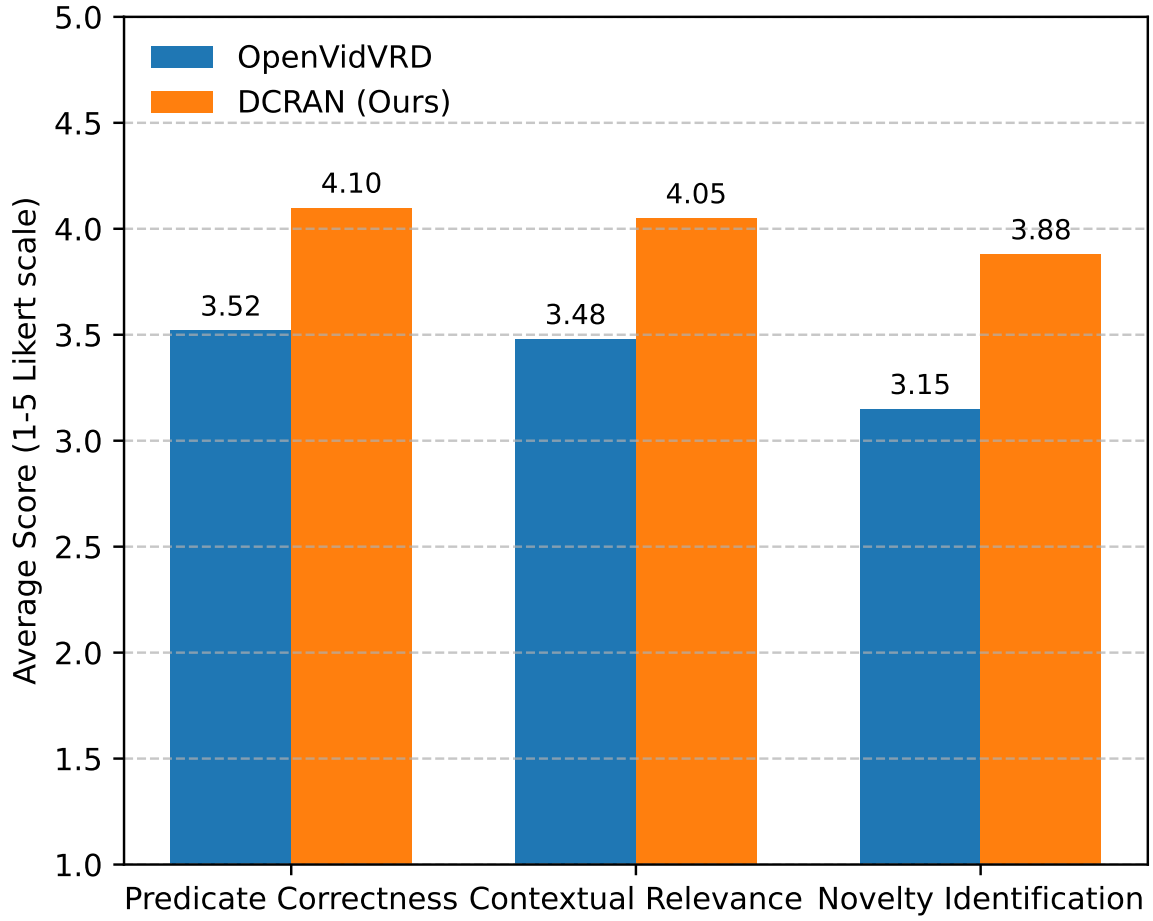

Figure 5. Average human evaluation scores (1-5 Likert scale, higher is better) for novel relation predictions on the VidOR dataset. Results demonstrate DCRAN's superior ability to capture correct, contextually relevant, and novel relations. The data presented is illustrative.

The human evaluation results corroborate our quantitative findings. As depicted in Figure 5, DCRAN consistently received higher average scores across all criteria. Specifically, its superior performance in

 doi:10.20944/preprints202511.1974.v1

"Predicate Correctness" and "Contextual Relevance" highlights the effectiveness of the Spatio-Temporal Gating Module and the Dynamic Relational Prompting module in understanding the dynamic nature of video interactions. The most significant improvement was observed in "Novelty Identification," where DCRAN achieved an average score of 3.88 compared to OpenVidVRD's 3.15. This substantial gain directly validates the design principles of DCRAN, particularly the Multi-Granular Semantic Alignment module and the dynamic prompting mechanism, which enable the model to infer and articulate unseen relationships with greater semantic accuracy and contextual appropriateness. Qualitatively, DCRAN was observed to predict more nuanced and specific predicates for novel interactions, such as "person *gesturing at* screen" instead of a generic "person *looking at* screen" or "car *skidding on* road" instead of "car *moving on* road," demonstrating its fine-grained understanding of dynamic video content."

5. Conclusions

In this paper, we introduced the Dynamic Contextual Relational Alignment Network (DCRAN), a novel end-to-end framework designed for open-vocabulary video visual relation detection, addressing the challenge of understanding dynamic relationships and generalizing to unseen predicate categories. DCRAN integrates a Contextual Tracklet Feature Extractor with a Spatio-Temporal Gating Module (STGM) for refining VLM features, a Dynamic Relational Prompting (DRP) Module for generating context-aware predicate prompt vectors, and a Multi-Granular Semantic Alignment (MGSA) Module with a Hierarchical Contrastive Loss for robust alignment of objects and dynamic relations. Comprehensive experiments on VidVRD and VidOR datasets demonstrated DCRAN's superior, state-of-the-art performance across all metrics, particularly exhibiting significant generalization capabilities in demanding Novel-split settings. Ablation studies confirmed the critical contributions of each module, while analysis highlighted the semantic alignment of DRP and DCRAN's efficiency due to its frozen VLM backbone. Qualitative and human evaluations further validated its ability to identify correct and contextually relevant novel relations. We believe DCRAN's principles of dynamic context integration and multi-granular semantic alignment hold significant promise for advancing other open-vocabulary video understanding tasks, paving the way for more intelligent and adaptable AI systems.

References

1. Liu, X.; Huang, H.; Shi, G.; Wang, B. Dynamic Prefix-Tuning for Generative Template-based Event Extraction. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 5216–5228. https://doi.org/10.18653/v1/2022.acl-long.358.
2. Mishra, S.; Khashabi, D.; Baral, C.; Hajishirzi, H. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 3470–3487. https://doi.org/10.18653/v1/2022.acl-long.244.
3. Zhou, Y.; Shen, J.; Cheng, Y. Weak to strong generalization for large language models with multi-capabilities. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
4. Hu, D.; Wei, L.; Huai, X. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 7042–7052. https://doi.org/10.18653/v1/2021.acl-long.547.
5. Ding, N.; Chen, Y.; Xu, B.; Qin, Y.; Hu, S.; Liu, Z.; Sun, M.; Zhou, B. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 3029–3051. https://doi.org/10.18653/v1/2023.emnlp-main.183.
6. Xu, H.; Yan, M.; Li, C.; Bi, B.; Huang, S.; Xiao, W.; Huang, F. E2E-VLP: End-to-End Vision-Language Pre-training Enhanced by Visual Learning. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 503–513. https://doi.org/10.18653/v1/2021.acl-long.42.

 doi:10.20944/preprints202511.1974.v1

7. Zhou, Y.; Li, X.; Wang, Q.; Shen, J. Visual In-Context Learning for Large Vision-Language Models. In Proceedings of the Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. Association for Computational Linguistics, 2024, pp. 15890–15902.
8. Gao, K.; Chen, L.; Zhang, H.; Xiao, J.; Sun, Q. Compositional Prompt Tuning with Motion Cues for Open-vocabulary Video Relation Detection. *arXiv preprint arXiv:2302.00268v1* **2023**.
9. Macé, Q.; Loison, A.; Faysse, M. ViDoRe Benchmark V2: Raising the Bar for Visual Retrieval. *arXiv preprint arXiv:2505.17166v2* **2025**.
10. Chen, Z.; Chen, W.; Smiley, C.; Shah, S.; Borova, I.; Langdon, D.; Moussa, R.; Beane, M.; Huang, T.H.; Routledge, B.; et al. FinQA: A Dataset of Numerical Reasoning over Financial Data. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 3697–3711. https://doi.org/10.18653/v1/2021.emnlp-main.300.
11. Gao, J.; Sun, X.; Xu, M.; Zhou, X.; Ghanem, B. Relation-aware Video Reading Comprehension for Temporal Language Grounding. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 3978–3988. https://doi.org/10.18653/v1/2021.emnlp-main.324.
12. Cao, M.; Chen, L.; Shou, M.Z.; Zhang, C.; Zou, Y. On Pursuit of Designing Multi-modal Transformer for Video Grounding. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 9810–9823. https://doi.org/10.18653/v1/2021.emnlp-main.773.
13. Wang, Z.; Wen, J.; Han, Y. EP-SAM: An Edge-Detection Prompt SAM Based Efficient Framework for Ultra-Low Light Video Segmentation. In Proceedings of the ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025, pp. 1–5.
14. Lin, Z.; Zhang, Q.; Tian, Z.; Yu, P.; Lan, J. DPL-SLAM: enhancing dynamic point-line SLAM through dense semantic methods. *IEEE Sensors Journal* **2024**, *24*, 14596–14607.
15. Lin, Z.; Tian, Z.; Zhang, Q.; Zhuang, H.; Lan, J. Enhanced visual slam for collision-free driving with lightweight autonomous cars. *Sensors* **2024**, *24*, 6258.
16. Li, Q.; Tian, Z.; Wang, X.; Yang, J.; Lin, Z. Efficient and Safe Planner for Automated Driving on Ramps Considering Unsatisfication. *arXiv preprint arXiv:2504.15320* **2025**.
17. Long, Q.; Wang, M.; Li, L. Generative Imagination Elevates Machine Translation. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 5738–5748. https://doi.org/10.18653/v1/2021.naacl-main.457.
18. Wang, Z.; Jiang, W.; Wu, W.; Wang, S. Reconstruction of complex network from time series data based on graph attention network and Gumbel Softmax. *International Journal of Modern Physics C* **2023**, *34*, 2350057.
19. Vidgen, B.; Thrush, T.; Waseem, Z.; Kiela, D. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 1667–1682. https://doi.org/10.18653/v1/2021.acl-long.132.
20. Xu, H.; Ghosh, G.; Huang, P.Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; Feichtenhofer, C. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 6787–6800. https://doi.org/10.18653/v1/2021.emnlp-main.544.
21. Huang, P.Y.; Patrick, M.; Hu, J.; Neubig, G.; Metze, F.; Hauptmann, A. Multilingual Multimodal Pre-training for Zero-Shot Cross-Lingual Transfer of Vision-Language Models. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 2443–2459. https://doi.org/10.18653/v1/2021.naacl-main.195.
22. Wang, Z.; Xiong, Y.; Horowitz, R.; Wang, Y.; Han, Y. Hybrid Perception and Equivariant Diffusion for Robust Multi-Node Rebar Tying. In Proceedings of the 2025 IEEE 21st International Conference on Automation Science and Engineering (CASE). IEEE, 2025, pp. 3164–3171.
23. Wang, P.; Zhu, Z.; Feng, Z. Virtual Back-EMF Injection-based Online Full-Parameter Estimation of DTP-SPMSMs Under Sensorless Control. *IEEE Transactions on Transportation Electrification* **2025**.
24. Wang, P.; Zhu, Z.Q.; Feng, Z. Novel Virtual Active Flux Injection-Based Position Error Adaptive Correction of Dual Three-Phase IPMSMs Under Sensorless Control. *IEEE Transactions on Transportation Electrification* **2025**.

25. Wang, P.; Zhu, Z.; Liang, D. Improved position-offset based online parameter estimation of PMSMs under constant and variable speed operations. *IEEE Transactions on Energy Conversion* **2024**, *39*, 1325–1340.
26. Pfeiffer, J.; Vulić, I.; Gurevych, I.; Ruder, S. UNKs Everywhere: Adapting Multilingual Language Models to New Scripts. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 10186–10203. https://doi.org/10.18653/v1/2021.emnlp-main.800.
27. Bao, S.; He, H.; Wang, F.; Wu, H.; Wang, H.; Wu, W.; Guo, Z.; Liu, Z.; Xu, X. PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2513–2525. https://doi.org/10.18653/v1/2021.findings-acl.222.
28. Chen, C.Y.; Li, C.T. ZS-BERT: Towards Zero-Shot Relation Extraction with Attribute Representation Learning. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 3470–3479. https://doi.org/10.18653/v1/2021.naacl-main.272.
29. Ling, Y.; Yu, J.; Xia, R. Vision-Language Pre-Training for Multimodal Aspect-Based Sentiment Analysis. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 2149–2159. https://doi.org/10.18653/v1/2022.acl-long.152.
30. Wu, Y.; Lin, Z.; Zhao, Y.; Qin, B.; Zhu, L.N. A Text-Centered Shared-Private Framework via Cross-Modal Prediction for Multimodal Sentiment Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 4730–4738. https://doi.org/10.18653/v1/2021.findings-acl.417.
31. Shi, Z.; Zhou, Y. Topic-selective graph network for topic-focused summarization. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2023, pp. 247–259.
32. Wang, F.; Shi, Z.; Wang, B.; Wang, N.; Xiao, H. Readerlm-v2: Small language model for HTML to markdown and JSON. *arXiv preprint arXiv:2503.01151* **2025**.
33. Ostendorff, M.; Rethmeier, N.; Augenstein, I.; Gipp, B.; Rehm, G. Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 11670–11688. https://doi.org/10.18653/v1/2022.emnlp-main.802.
34. Shi, Z.; Zhou, Y.; Li, J.; Jin, Y.; Li, Y.; He, D.; Liu, F.; Alharbi, S.; Yu, J.; Zhang, M. Safety alignment via constrained knowledge unlearning. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 25515–25529.
35. Hui, J.; Tang, K.; Zhou, Y.; Cui, X.; Han, Q. The causal impact of gut microbiota and metabolites on myopia and pathological myopia: a mediation Mendelian randomization study. *Scientific Reports* **2025**, *15*, 12928.
36. Wang, J.; Cui, X. Multi-omics Mendelian Randomization Reveals Immunometabolic Signatures of the Gut Microbiota in Optic Neuritis and the Potential Therapeutic Role of Vitamin B6. *Molecular Neurobiology* **2025**, pp. 1–12.
37. Cui, X.; Liang, T.; Ji, X.; Shao, Y.; Zhao, P.; Li, X. LINC00488 induces tumorigenicity in retinoblastoma by regulating microRNA-30a-5p/EPHB2 Axis. *Ocular Immunology and Inflammation* **2023**, *31*, 506–514.
38. Ren, L. AI-Powered Financial Insights: Using Large Language Models to Improve Government Decision-Making and Policy Execution. *Journal of Industrial Engineering and Applied Science* **2025**, *3*, 21–26.
39. Ren, L.; et al. Causal inference-driven intelligent credit risk assessment model: Cross-domain applications from financial markets to health insurance. *Academic Journal of Computing & Information Science* **2025**, *8*, 8–14.
40. Ren, L. Causal Modeling for Fraud Detection: Enhancing Financial Security with Interpretable AI. *European Journal of Business, Economics & Management* **2025**, *1*, 94–104.
41. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning. PmLR, 2021, pp. 8748–8763.