TD^2-Net: Toward Denoising and Debiasing for Dynamic Scene Graph Generation

Xin Lin[1]*, Chong Shi[1], Yibing Zhan[2], Zuopeng Yang[1]*, Yaqi Wu[1], Dacheng Tao[3]

[1] Guangzhou University
[2] JD Explore Academy
[3] The University of Sydney
linxin94@gzhu.edu.cn, {shichong, winnerwu}@e.gzhu.edu.cn, zybjy@mail.ustc.edu.cn
{yzpeng44, dacheng.tao}@gmail.com

Abstract

Dynamic scene graph generation (SGG) focuses on detecting objects in a video and determining their pairwise relationships. Existing dynamic SGG methods usually suffer from several issues, including 1) Contextual noise, as some frames might contain occluded and blurred objects. 2) Label bias, primarily due to the high imbalance between a few positive relationship samples and numerous negative ones. Additionally, the distribution of relationships exhibits a long-tailed pattern. To address the above problems, in this paper, we introduce a network named TD^2-Net that aims at denoising and debiasing for dynamic SGG. Specifically, we first propose a denoising spatio-temporal transformer module that enhances object representation with robust contextual information. This is achieved by designing a differentiable Top-K object selector that utilizes the gumbel-softmax sampling strategy to select the relevant neighborhood for each object. Second, we introduce an asymmetrical reweighting loss to relieve the issue of label bias. This loss function integrates asymmetry focusing factors and the volume of samples to adjust the weights assigned to individual samples. Systematic experimental results demonstrate the superiority of our proposed TD^2-Net over existing state-of-the-art approaches on Action Genome databases. In more detail, TD^2-Net outperforms the second-best competitors by 12.7 % on mean-Recall@10 for predicate classification.

Introduction

A scene graph is a graph-structured representation that uses nodes to represent objects and edges to represent relationships in an image. It provides a practical approach for scene understanding, serving as a bridge between visual and language modalities, and is widely applied in various fields such as image captioning (Yang et al. 2020), image retrieval (Johnson et al. 2015), and visual question answering (Yang et al. 2020). While the development of scene graph generation (SGG) of images has been satisfactory, research on dynamic scene graph generation of videos is still in its infancy.

Dynamic SGG aims to detect the objects for each frame and the relationships among them. Such a detailed and structured video understanding is akin to how humans perceive real-world activities. Existing approaches (Cong et al. 2021; Nag et al. 2023; Feng et al. 2023) in dynamic SGG predominantly utilize transformers (Vaswani et al. 2017) for context modeling to acquire spatial-temporal information of objects or predicates. Furthermore, some methods employ strategies such as temporal prior inference (Wang et al. 2022), uncertainty-aware learning, and memory-guided training (Nag et al. 2023) to achieve unbiased dynamic SGG.

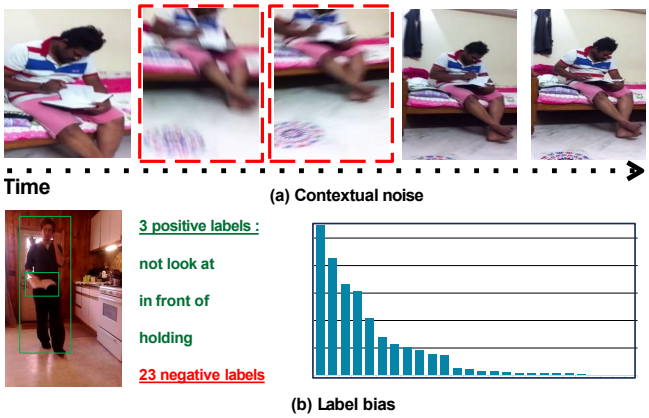

Figure 1: (a) Contextual noise. A significant proportion of objects may be occluded or affected by camera motion blur. (b) Label bias. As shown in the left example of two objects, the quantity of positive relationship labels is significantly less than that of negative ones, causing a negative-positive imbalance. Furthermore, as shown in the right tabular, the distribution of relationships exhibits a long-tailed trend.

However, as shown in Figure 1, the current dynamic SGG still faces two main problems. Firstly, there is the issue of contextual noise, as videos consist of noisy and correlated sequences of frames (Buch et al. 2022). A significant fraction of objects may be occluded or affected by camera motion blur, as depicted in Figure 1(a). When acquiring contextual information for objects or relationships, the irrelevant objects may introduce redundant or erroneous information, leading to unnecessary computational overhead and reducing the accuracy of the model's predictions. Secondly, there is the problem of label bias. As depicted in Figure 1(b), the predicate classes in standard datasets exhibit two types

*Corresponding Author

of imbalances, including positive-negative imbalance and head-tail imbalance. The former may result in the potential underestimation of the gradient of positive labels during training, as negative labels are often much more numerous than positive labels. The long-tailed distribution of positive labels may lead to the model's inability to recognize rare positive samples effectively, thereby reducing the accuracy and diversity of the model's predictions.

To address the abovementioned issues, we propose a network named (TD^2-Net) that aims at denoising and debiasing for dynamic SGG. Firstly, we introduce the denoising spatio-temporal transformer (D-Trans) module to tackle the contextual noise. Specifically, we achieve preliminary object matching by considering objects' appearance and spatial location consistency across frames. Furthermore, to eliminate noisy spatio-temporal information caused by irrelevant objects, we introduce the Gumbel-Softmax sampling strategy (Jang, Gu, and Poole 2016). This ensures that each object only aligns with the most relevant neighboring objects within a video, selecting more appropriate contextual information. Afterward, we enhance the object representations by aggregating contextual information that have been selected.

Secondly, we introduce the asymmetrical reweighting loss (AR-Loss) to address the label bias issue in relationship prediction. To tackle the positive-negative imbalance problem, we utilize different values of focusing factors for positive and negative samples, controlling their contribution to the loss function. Specifically, we set the focusing factor of positive samples to be higher than that of negative ones, leading the model to place more emphasis on positive samples. Moreover, we incorporate the concept of the effective number of samples, as described by (Cui et al. 2019), to mitigate the problem of head-tail imbalance. This adjustment allows the model to learn meaningful features from positive samples, despite their rarity.

In summary, the contributions of this study are twofold: (1) D-Trans for enhancing object features with denoising contextual information; (2) AR-Loss to deal with both positive-negative imbalance and head-tail imbalance in relationship prediction. The efficacy of the proposed TD^2-Net is systematically evaluated on the video scene graph generation benchmark dataset. Experimental results show that our TD^2-Net consistently outperforms state-of-the-art methods.

Related Work

Image Scene Graph Generation

Existing works in image SGG (ImgSGG) typically focus on context modeling or tackling the class imbalance problem (*i.e.*, the long-tailed distribution). Several context modeling strategies (Lin et al. 2020; Li et al. 2021), have been proposed to learn discriminative object representation by exploring various message passing mechanisms. Specifically, Lin *et al.* build a heterophily-aware message-passing scheme to distinguish the heterophily and homophily between objects/relationships (Lin et al. 2022a). Tang *et al.* proposes a dynamic tree structure to capture task-specific visual contexts (Tang et al. 2019). To handle the class imbalance issue, Lin *et al.* propose a group diversity enhancement module to relieve low diversity in relationship predictions (Lin et al. 2022b). Zheng *et al.* relieves the ambiguous entity-predicate matching caused by the predicate's semantic overlap by prototype regularization (Zheng et al. 2023). However, modeling each frame in long videos can result in high computational complexity and redundant information. Current ImgSGG methods prioritize spatial over temporal information, thus missing inter-frame correlations. Furthermore, unlike ImgSGG, which focuses on a single-label biased problem, video SGG or dynamic SGG deals with a more complicated multi-label biased problem.

Video Scene Graph Generation

Compared with ImgSGG, video SGG (VidSGG) or dynamic SGG is more challenging because it needs to consider the spatio-temporal context in adjacent frames. Existing VidSGG methods can be roughly divided into two categories, including tracklet-based and frame-based approaches. Specifically, for the tracklet-based method, each graph node is an object tracklet within a video segment (Shang et al. 2017, 2019) or the entire video (Liu et al. 2020; Gao et al. 2022, 2023b). For the frame-based method, each graph node is an object box, similar to ImgSGG, but the visual relation triplets are dynamic throughout the entire video sequence (Ji et al. 2020; Feng et al. 2023; Cong et al. 2021; Li, Yang, and Xu 2022). Additionally, researchers have begun to focus on addressing the issue of biased prediction in VidSGG. In more detail, Wang *et al.* utilizes temporal prior knowledge as an inductive bias to generate dynamic scene graph (Wang et al. 2022). Nag *et al.* learns to synthesize unbiased relationship representations using memory-guided training and attenuates the predictive uncertainty of visual relations using a Gaussian Mixture Model (Nag et al. 2023). However, current VidSGG methods still face two main problems. Firstly, there is the issue of contextual noise caused by certain frames in a video that might contain occluded or blurred objects. Secondly, the predicate classes exhibit two types of imbalances, including positive-negative imbalance and head-tail imbalance.

Method

The framework of the proposed TD^2-Net is illustrated in Figure 2. We employ Faster R-CNN (Ren et al. 2015) to obtain object proposals for each video frame. We adopt the same way as (Cong et al. 2021) to obtain the feature for each proposal. The appearance features, spatial feature, and classification score vector for the i-th object is denoted as $\boldsymbol{x}_i$, $\boldsymbol{b}_i$, and $\boldsymbol{p}_i$, respectively. We further extract features from the union box of one pair of proposal i and j, denoted as x_{ij}. To achieve denoising and debiasing for dynamic SGG, we have made two contributions in this work. Firstly, a denoising spatio-temporal transformer (D-Trans) module is introduced to address the issue of context noise. Secondly, an asymmetrical reweighting loss (AR-Loss) is introduced to tackle the label bias problem. In the below, we will describe these two components sequentially.

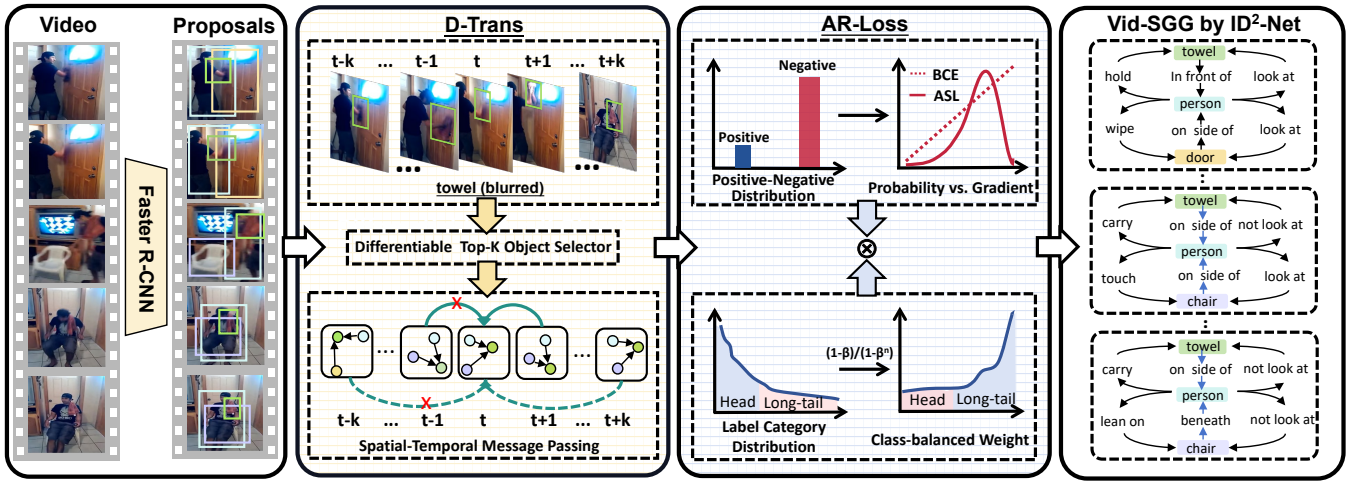

Figure 2: The framework of TD^2-Net. TD^2-Net adopts Faster-RCNN to generate initial object proposals for each RGB frame in a video. It includes two new modules for dynamic scene graph generation: (1) a novel transformer module named D-Trans that enhances object feature with robust contextual information (2) a new loss function named AR-Loss that takes into account both positive-negative imbalance and head-tail imbalance in relationship prediction.

Denosing Spatio-Temporal Transformer

Existing works on VidSGG (Gao et al. 2021; Nag et al. 2023; Feng et al. 2023) typically gather contextual information by tracking or matching objects in sequential frames. However, these methods are prone to being affected by contextual noise, leading to some spurious and redundant object correlations, which reduces the accuracy of model predictions and results in unnecessary resource consumption. To alleviate this issue, we propose the D-Trans module, which includes a differentiable Top-K object selector and spatial-temporal message passing.

Differentiable Top-K Object Selector. Inspired by the video graph representation method proposed in (Xiao et al. 2022), we employ a score function based on objects' appearance and spatial location to ensure temporal consistency across frames. Specifically, the matching score for two objects i and j in different frames can be expressed as follows:

$$g_{ij} = \psi\left(\boldsymbol{p}_i, \boldsymbol{p}_j\right) + \mathrm{IoU}\left(\boldsymbol{b}_i, \boldsymbol{b}_j\right), \tag{1}$$

where ψ and IoU denote the cosine similarity and intersection-over-union functions, respectively. Detected objects in consecutive frames are linked to the target objects by greedily maximizing Eq. (1) frame by frame. By aligning objects within a video, we ensure the consistency of the object representations for the graphs constructed at different frames. By gathering all the appearance features of the aligned objects for the i-th object, we can obtain its neighborhood feature matrix $\boldsymbol{Z}_i$.

Compared with the object matching strategy in existing frame-based dynamic SGG works, Eq. (1) has two advantages. First, compared with utilizing the predicted object label (Nag et al. 2023), it utilizes the score vector of objects which can better reflect the uncertainty of the prediction. Besides, the IoU function can ensure temporal consistency. Second, compared with utilizing Hungarian matching (Feng et al. 2023), it enjoys a low computational complexity.

However, as the number of video frames increases in long videos, the number of objects associated with a target object also increases. In order to dynamically eliminate unreliable spatio-temporal information caused by irrelevant objects, inspired by the method (Gao et al. 2023a) of using a segment selector to extract the relevant frames for the given question, we propose a differentiable Top-K selector to choose the most relevant contextual information from neighboring objects for the target object, which is defined as follows:

$$\boldsymbol{F}_i = \sigma(\underset{\text{Top-K}}{\text{selector}}(\text{softmax}(\frac{\boldsymbol{x}_i\boldsymbol{K}^T}{\sqrt{d_k}}, \boldsymbol{V}))), \tag{2}$$

where σ represents the function that flattens the matrix into a vector by row. $\boldsymbol{K}=\boldsymbol{V}=\boldsymbol{Z}_i+\boldsymbol{E}_i$, d_k denotes the dimension of $\boldsymbol{K}$, the $\boldsymbol{E}_i$ represents positional encodings (Vaswani et al. 2017) that provide temporal location information for each aligned object relative to the i-th object. In more detail, the Top-K selection can be implemented by extending the gumbel-softmax trick (Jang, Gu, and Poole 2016) or based on optimal transport formulations (Xie et al. 2020) for ranking and sorting. In this paper, we conduct gumbel-softmax sampling K times with replacement to achieve Top-K selection. Note that we sample the objects with replacement, as certain objects may only exist in a few frames. In such cases, we aim to guide the model learns to select the most related object by re-sampling it instead of forcing it to select irrelevant objects, as sampling without replacement does.

Spatial-Temporal Message Passing. Given the selected neighborhood for each object, we further stack N standard multi-head attention (MHA) layers (Vaswani et al. 2017) to enhance the object representations by aggregating information from other aligned objects from the adjacent frames within a video:

$$\boldsymbol{X}^{'} = \mathrm{MHA}^{(N)}_{\mathrm{temporal}}(\Phi^t_q(\boldsymbol{X}+\boldsymbol{E}), \Phi^t_k(\boldsymbol{F}), \Phi^t_v(\boldsymbol{F})), \quad (3)$$

where Φ^t_q, Φ^t_k, and Φ^t_v are linear transformation. In addition, we apply M MHA layers to reason over the object spatial interactions as follows:

$$\hat{\boldsymbol{X}} = \mathrm{MHA}^{(M)}_{\mathrm{spatial}}(\Phi^s_q(\boldsymbol{X}^{'}_f), \Phi^s_k(\boldsymbol{X}^{'}_f), \Phi^s_v(\boldsymbol{X}^{'}_f)), \quad (4)$$

where Φ^s_q, Φ^s_k, and Φ^s_v denote linear transformation, $\boldsymbol{X}^{'}_f$ denotes the representations of objects from the same frame.

In summary, our motivation for the D-Trans module is that it models the change of object behaviors and thus infers dynamic actions. Also, it is helpful in improving the objects' appearance feature in cases where the object at certain frames suffers from motion blur or partial occlusion.

Asymmetrical Reweighting Loss

After obtaining the refined object representation by D-Trans, we further propose an asymmetrical reweighting loss (AR-Loss) to mitigate the issue of label bias in relationship prediction. We adopt the same approach as in STTran (Cong et al. 2021) to obtain the classification score vector of the relationship between two objects i and j as follows:

$$\boldsymbol{p}_{ij} = \phi(\mathrm{RTrans}([\boldsymbol{W}_s\hat{\boldsymbol{x}}_i, \boldsymbol{W}_o\hat{\boldsymbol{x}}_j, \boldsymbol{x}_{ij}, \boldsymbol{c}_i, \boldsymbol{c}_j])), \quad (5)$$

where RTrans denotes that we utilize the same structure as our proposed spatial-temporal message passing module to refine the relationship feature. ϕ indicates the classification operation. $\boldsymbol{c}_i$, $\boldsymbol{c}_j$ denote the semantic embedding of the i-th object and j-th object. $\boldsymbol{W}_s$, $\boldsymbol{W}_o$ are projection matrices for fusion, $[,]$ represents the concatenation operation.

Binary cross-entropy loss (Nag et al. 2023; Wang et al. 2022) and multi-label margin loss (Cong et al. 2021) have been widely adopted for optimization in previous dynamic SGG models. However, the aforementioned loss functions consider each sample to be equally significant and assign them the same weight, which may not be suitable to alleviate the issue of label bias. To address this problem, we first revisit the focal loss (Lin et al. 2017), which is a traditional solution to mitigate the positive-negative imbalance issue. It adjusts the loss contribution of easy and hard samples, which reduces the influence of the majority of negative samples:

$$\mathcal{L}_{fl}\left(\boldsymbol{p}_{ij}\right) = \begin{cases} \mathcal{L}^+_{fl} = \left(1-\boldsymbol{p}_{ij}\right)^{\gamma}\log\left(\boldsymbol{p}_{ij}\right), & \text{if } y=1 \\ \mathcal{L}^-_{fl} = \boldsymbol{p}^{\gamma}_{ij}\log(1-\boldsymbol{p}_{ij}), & \text{if } y=0 \end{cases} \quad (6)$$

where y is the ground-truth label, and $\mathcal{L}^+_{fl}$ and $\mathcal{L}^-_{fl}$ represent the loss function of positive and negative samples, respectively. γ denotes the focusing parameter. By setting $\gamma > 0$, It reduces the impact of easy negatives on the loss function. However, the focal loss may not sufficiently address the following issues:

- **Positive-Negative Imbalance.** Most object pairs contain fewer positive labels and more negative labels on average. A High value of γ in Eq. (6) sufficiently down-weights the contribution from easy negatives but may eliminate the gradients from the tail positive samples.
- **Head-Tail Imbalance.** Due to the long-tailed distribution of the datasets, the head-tail imbalance exists in the positive samples. This imbalance between different categories of positive samples may lead to the model failing to recognize rare positive samples.

Thus, we decouple the focusing levels of the positive and negative samples to alleviate the positive-negative imbalance. Specifically, we set γ^+ and γ^- to be the positive and negative focusing parameters, respectively. Furthermore, to mitigate the impact of head-tail imbalance in the positive samples, inspired by (Cui et al. 2019), which adjusts sample weights using the effective number of samples for each class, we defined ω_{cb} as follows to adjust the weights assigned to individual sample:

$$\omega_{cb} = \frac{1-\beta}{1-\beta^{n_{\hat{y}}}}, \quad (7)$$

where $n_{\hat{y}}$ is the number of samples of the ground-truth class $\hat{y}$ in the training set. A higher value of ω_{cb} for tail samples will increase their weight, encouraging the model to pay more attention to the positive tail samples and vice versa. The hyper-parameter $\beta \in [0, 1)$ controls the rate at which the weight grows as $n_{\hat{y}}$ increases.

After applying the asymmetric focusing factors γ^+, γ^-, and the effective number of samples ω_{cb} into our AR-Loss, we obtain the loss function as follows:

$$\mathcal{L}_{ar}(\boldsymbol{p}_{ij}) = \begin{cases} \mathcal{L}^+_{ar} = \omega_{cb}\left(1-\boldsymbol{p}_{ij}\right)^{\gamma^+}\log\left(\boldsymbol{p}_{ij}\right), & \text{if } y=1 \\ \mathcal{L}^-_{ar} = \boldsymbol{p}^{\gamma^-}_{ij}\log(1-\boldsymbol{p}_{ij}). & \text{if } y=0 \end{cases} \quad (8)$$

Note that $\gamma^+ = \gamma^- = 0$ and ω_{cb}=1 yields binary cross-entropy. Since we are interested in emphasizing the contribution of positive samples, we set $\gamma^- \geq \gamma^+$. We achieve better control over the contribution of positive and negative samples through Eq. (8), which assists the network in learning meaningful features from positive samples, despite their rarity. Thus, AR-loss can simultaneously address the positive-negative imbalance and head-tail imbalance.

VidSGG by TD^2-Net

During training, the overall loss function $\mathcal{L}$ for TD^2-Net can be expressed as follows:

$$\mathcal{L} = \mathcal{L}_{obj} + \mathcal{L}_{ar}, \quad (9)$$

where $\mathcal{L}_{obj}$ denotes the cross entropy loss for object classification.

During testing, the score of each relationship triplet <subject-predicate-object> is computed as:

$$s_{rel} = s_{sub} * s_p * s_{obj}, \quad (10)$$

where s_{sub}, s_p, s_{obj} are the predicted score of subject, predicate, and object, respectively.

Experiments

Dataset and Evaluation Setting

Dataset. Our experiments are conducted on the AG dataset (Ji et al. 2020), which is the benchmark dataset of

Method	With Constraint						No Constraint					
	PredCLS		SGCLS		SGDET		PredCLS		SGCLS		SGDET	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
VRD (2016)	51.7	54.7	32.4	33.3	19.2	26.0	59.6	99.2	39.2	52.6	19.1	40.5
Motif Freq(2018)	62.4	65.1	40.8	41.9	23.7	33.3	73.4	99.6	50.4	64.2	22.8	46.4
MSDN (2017)	65.5	68.5	43.9	45.1	24.1	34.5	74.9	99.0	51.2	65.0	23.1	46.5
VCTREE (2019)	66.0	69.3	44.1	45.3	24.4	34.7	75.5	99.3	52.4	65.1	23.9	46.8
RelDN (2019)	66.3	69.5	44.3	45.4	24.5	34.9	75.7	99.0	52.9	65.1	24.1	46.8
GPS-Net (2020)	66.8	69.9	45.3	46.5	24.7	35.1	76.0	99.5	53.6	66.0	24.4	47.3
STTran (2021)	68.6	71.8	46.4	47.5	25.2	37.0	77.9	99.1	54.0	66.4	24.6	48.8
TPI (2022)	69.7	72.6	47.2	48.3	26.2	**37.4**	-	-	-	-	-	-
TEMP (2023)	68.8	71.5	47.2	48.3	28.1	34.9	80.4	99.4	56.3	67.9	29.8	46.4
TD^2-Net (P)	**70.1**	**73.1**	**51.1**	**52.1**	**28.7**	37.1	**81.7**	**99.8**	**57.2**	**69.8**	**30.5**	**49.3**
TD^2-Net	67.8	70.8	49.1	50.2	27.2	36.7	78.2	99.2	55.1	67.3	28.1	48.4

Table 1: Comparisons with state-of-the-art on the Action Genome dataset. The same object detector is used in all baselines for fair comparison. TD^2-Net (P) indicate that we set ω_{cb} as 1 in AR-Loss. The best methods are marked according to formats under each setting.

dynamic scene graph generation. AG is built upon the Charades dataset (Sigurdsson et al. 2016) and provides frame-level scene graph labels with a total of 234,253 frames in 9,848 video clips. In AG, there are 36 types of entities and 26 types of relations in the label annotations. Such 26 types of relations are divided into three classes (*i.e.,* attention, spatial, and contacting relations). The attention relations are used to describe if a person is looking at an object or not. The spatial relations specify the relative position. The contacting relations represent different ways of contacting in particular.

Evaluation Setting. We use the same data and evaluation metrics that have been widely adopted in recent works (Ji et al. 2020; Cong et al. 2021; Nag et al. 2023; Wang et al. 2022). Specifically, We make the evaluation of TD^2-Net on the AG dataset under three conventional tasks below: (1) Predicate Classification (PredCLS): Given the ground-truth object bounding boxes and categories, the model needs to predict predicate categories; (2) Scene Graph Classification (SGCLS): Given the ground-truth bounding boxes of objects, the model needs to predict both the object and relationship categories; (3) Scene Graph Detection (SGDET): Given an image, the model detects object and predict relationship categories between each pair of objects. All algorithms are evaluated using the Recall@K (R@K) and mean-Recall@K (mR@K) metrics, for K=$[10, 50]$. Evaluation is conducted under two setups: **With Constraints** and **No Constraints** to make a fair and sufficient comparison with baselines. In more detail, **With Constraints** is the most stringent since it only chooses one predicate for each entity pair. **No Constraints** allows multiple predictions of relations for each entity pair, taking top 100 predicates for all pairs in a single frame.

Implementation Details. To ensure compatibility with previous state-of-the-art architectures, we follow STTran (Cong et al. 2021) and use Faster R-CNN (Ren et al. 2015) based on ResNet-101 (He et al. 2016) as the backbone for object detection. During training, we utilize the AdamW optimizer (Loshchilov and Hutter 2017) with an initial learning rate of $1e^{-5}$ and a batch size of 1. The model is trained for 10 epochs. Additionally, we apply gradient clipping, restricting the gradients to a maximum norm of 5. In the Eq. (3) and Eq. (4), we set parameters $M = N = 3$.

Comparisons with State-of-the-Art Methods

Table 1 shows that TD^2-Net (P) outperforms all state-of-the-art methods on various metrics. Specifically, TD^2-Net (P) outperforms the best ImgSGG method, named GPS-Net, by 3.3 %, 5.8%, and 4.0% at R@10 on PRECLS, SGCLS, and SGDET, respectively. Moreover, even when compared with the best unbiased dynamic SGG method TEMP, TD^2-Net (P) still demonstrates a performance improvement of 3.9% at R@10 on SGCLS task.

Due to the class imbalance problem in Action Genome, previous works usually achieve low performance for less frequent categories. Hence, we conduct an experiment utilizing the mR@K as evaluation metric (Nag et al. 2023) for all three SGG tasks under both **With Constraint** and **No Constraints** settings. We also relied on email communications with the authors of several papers on the mR values where the source code are not publicly available. As shown in Table 2, TD^2-Net shows a large absolute gain for the Mean Recall metric, which indicates that TD^2-Net has advantages in handling the class imbalance problem of dynamic SGG. In more detail, TD^2-Net outperform one very recent unbiased dynamic SGG method, named TEMP (Nag et al. 2023), by 12.7% at mR@10 on PREDCLS under with constraint setting. To illustrate this advantage more vividly, we present the R@10 improvement of each predicate category compared with STTran (Cong et al. 2021), and TRACE (Teng et al. 2021) for PREDCLS under **With Constraints** setting in Figure 3. These improvements are much larger for minority relationship categories. We owe this advantage to the power of the AR-Loss. Overall, TD^2-Net does not compromise Recall values and achieves comparable or better per-

Method	With Constraint						No Constraint					
	PredCLS		SGCLS		SGDET		PredCLS		SGCLS		SGDET	
	mR@10	mR@50	mR@10	mR@50	mR@10	mR@50	mR@10	mR@50	mR@10	mR@50	mR@10	mR@50
RelDN (2019)	6.2	6.2	3.4	3.4	3.3	3.3	31.2	75.5	18.6	42.6	7.5	37.7
TRACE (2021)	15.2	15.2	8.9	8.9	8.2	8.2	50.9	82.7	31.9	46.3	22.8	41.8
STTran (2021)	37.8	40.2	27.2	28.0	16.6	22.2	51.4	82.7	40.7	58.8	20.9	39.2
TPI (2022)	37.3	40.6	28.3	29.3	15.6	21.8	-	-	-	-	-	-
TEMP (2023)	42.9	46.3	34.0	35.2	18.5	23.7	61.5	98.0	48.3	66.4	24.7	43.7
TD^2-Net (P)	41.9	44.8	33.9	34.9	17.2	22.3	61.0	96.4	50.1	67.9	23.2	42.1
TD^2-Net	**54.2**	**57.1**	**40.9**	**42.0**	**20.4**	**26.1**	**68.3**	**98.2**	**51.4**	**69.1**	**27.9**	**46.3**

Table 2: Comparison on the mR@K metric between various methods across all the 26 relationship categories. Note that we adopt the same evaluation metric as TEMP (Nag et al. 2023). TD^2-Net (P) indicate that we set ω_{cb} as 1 in AR-Loss. The best methods under each setting are marked according to formats.

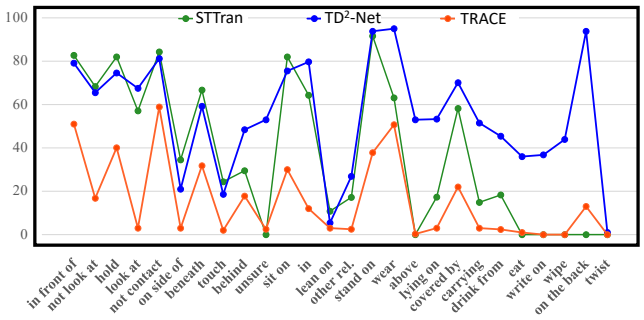

Figure 3: Comparative per class performance for PREDCLS task. Results are in terms of R@10 under With Constraint.

formance than the existing methods, which aim to achieve high Recall values without considering label bias issues.

Ablation Studies

We conduct four ablation studies to verify the effectiveness of our proposed methods. The results of the ablation studies are summarized in four tables: Table 3, Table 4, Table 5, and Table 6. It is worth noting that we have adopted TD^2-Net (P), which sets ω_{cb} as 1 in AR-Loss, in Table 4 and Table 5 to demonstrate the significant of each component of D-Trans.

Effectiveness of the Proposed Modules. We first perform an ablation study to justify the effectiveness of D-Trans and AR-Loss. The results are summarized in Table 3. Exp 1 in Table 3 shows the baseline performance based on STTran (Cong et al. 2021). To facilitate fair comparison, all the other settings remain the same as TD^2-Net. Exps 2-4 show that each module helps promote dynamic SGG performance. The best performance is achieved when both modules are involved. Note that D-Trans and AR-Loss are primarily designed to refine object and relationship representations, respectively. Therefore, D-Trans helps the model achieve outstanding SGCLS performance, which heavily depends on the object classification ability. Meanwhile, AR-Loss enables the model to achieve a significant performance gain on the PREDCLS task, mainly relying on relationship

Exp	Module		SGCLS		PredCLS	
	D-Trans	AR	R@10	mR@10	R@10	mR@10
1	-	-	46.4	27.2	68.6	37.8
2	✓	-	49.7	29.5	68.9	40.4
3	-	✓*	47.3	30.6	69.7	40.3
4	-	✓	45.7	37.3	67.3	53.1
5	✓	✓*	**51.1**	33.9	**70.1**	42.2
6	✓	✓	49.1	**40.9**	67.8	**54.2**

Table 3: Ablation studies. We consistently adopt the same object detection backbone as in (Cong et al. 2021). "✓*" denotes that we set ω_{cb} as 1 in AR-Loss.

	K	4	6	8	10
SGCLS	R@10	49.6	50.4	**51.1**	50.9
	R@20	50.7	51.4	**52.1**	51.9
	R@50	50.7	51.4	**52.1**	51.9

Table 4: Evaluation on the value of Top-K in Eq. (2).

prediction power.

Evaluation on Hyperparameters for D-Trans. we verify the impact of the hyperparameters of the D-Trans modules. As shown in Table 4, TD^2-Net (P) achieves the best performance when K is set to 8 in the differentiable Top-K frame selector. However, if this threshold is exceeded, the model's memory usage increases with decreased benefits. More details can be found in the supplemental file.

Design Choices for the D-Trans Module. In Table 5, we compare the performance of D-Trans with and without the two object-matching strategies described in Eq. (1) and Eq. (2). "w/o D-Trans" denotes that we remove the D-Trans module in TD^2-Net (P). "Full" represents that we simultaneously utilize Eq. (1) and Eq. (2) in differentiable Top-K object selector for TD^2-Net (P). Experimental results in Table 5 show that the two object-matching strategies consistently achieve better performance. Therefore, the effective-

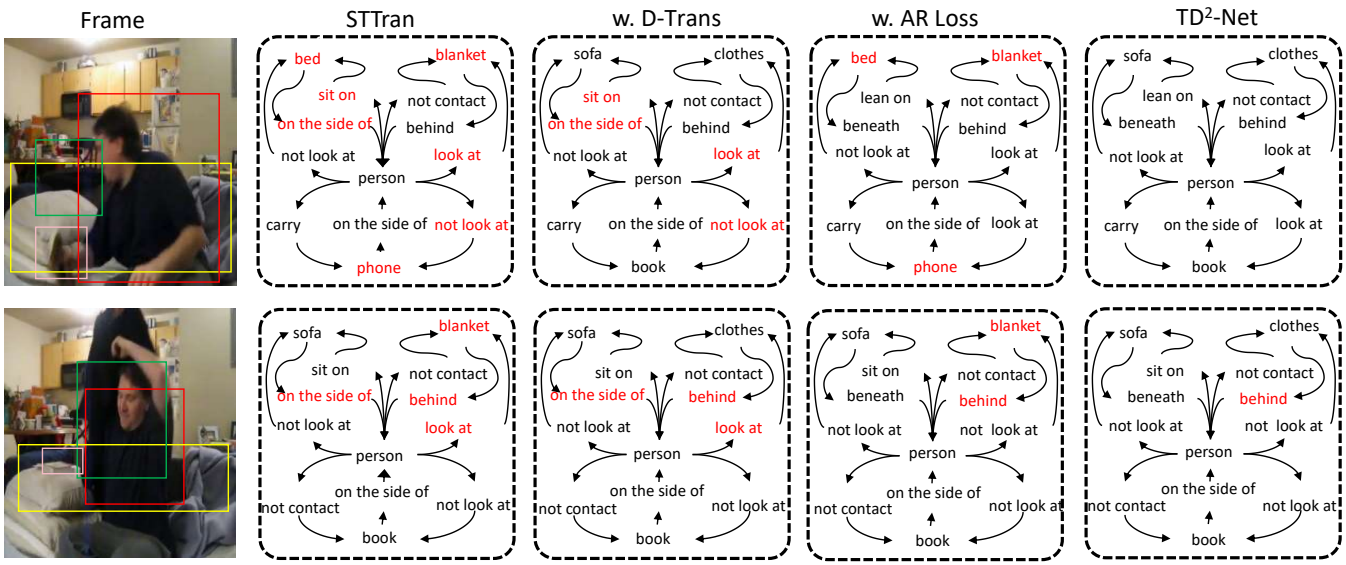

Figure 4: Qualitative comparisons between TD^2-Net and STTran (Cong et al. 2021). Specifically, we show the comparisons at R@100 in the SGCLS setting. The black color indicates correctly classified objects or predicates; the red indicates those that have been misclassified. Best viewed in color.

		w/o D-Trans	Eq. (1)	Eq. (2)	Full
SGCLS	R@10	47.3	49.1	50.7	**51.1**
	R@20	48.4	50.2	51.7	**52.1**
	R@50	48.4	50.2	51.7	**52.1**

Table 5: Design Choices for the D-Trans module.

		Focal	BCE	MLM	$\omega_{cb}=1$	AR
PREDCLS	R@10	69.4	69.2	69.0	**70.1**	67.8
	R@20	72.4	72.1	71.9	**73.1**	70.8
	R@50	72.4	72.1	71.9	**73.1**	70.8
	mR@10	42.1	41.2	40.4	41.9	**54.2**
	mR@20	44.3	43.4	42.7	44.8	**57.1**
	mR@50	44.3	43.4	42.7	44.8	**57.1**

Table 6: Evaluation on different choices of loss function.

ness of differentiable Top-K object selector is justified.

Comparisons between Four Loss Functions. We compare the performance of the Focal-Loss (Lin et al. 2017), AR-Loss (ω_{cb}=1), AR-Loss, and the other two kinds of loss functions which are widely utilized in the existing VidSGG methods: MLM-Loss (Cong et al. 2021), BCE-Loss (Nag et al. 2023). As shown in Table 6, AR-Loss (ω_{cb}=1) achieves the best performance in terms of R@K, as it effectively mitigates the issue of positive-negative imbalance. Additionally, when we focus on mR@K, AR-Loss outperforms the other methods. This can be attributed to the fact that AR-Loss effectively mitigates the issue of label bias.

Qualitative Evaluation. Figure 4 presents the qualitative results for the dynamic scene graph generation. The five columns from left to right are RGB frame, scene graph generated by STTran, STTran with D-Trans, STTran with AR-Loss, and scene graph generated by TD^2-Net, respectively. As can be seen from the third column of Figure 4, STTran with D-Trans produces superior object predictions compared to STTran for items such as "sofa", "book", and "clothes" that are challenging to identify from their proposals. Therefore, we owe this performance gain to the D-Trans module that utilizes robust context information to enhance the object's representation. In the fourth column of Figure 4, the improvements in predicates predictions, such as "lean on" and "behind", can be attributed to the contributions of AR-Loss. Furthermore, as illustrated in the rightmost column, TD^2-Net demonstrates superior overall performance.

Conclusion

In this paper, we propose a new model called TD^2-Net, which is designed to handle two critical issues in dynamic SGG: contextual noise and label bias. To address the contextual noise issue, we introduce a D-Trans module that uses a differentiable Top-K object selector to choose the most relevant neighborhood for each object. Then, we could enhance object representation with robust contextual information via spatio-temporal message passing. To mitigate the head-tail and positive-negative imbalance in relationship prediction, we introduce an AR-Loss which incorporates asymmetry focusing factors and sample volume to adjust the sample weights. Through extensive experiments on the Action Genome dataset, we demonstrate the effectiveness of our approach.

Acknowledgments

This work is supported by the Guangzhou basic and applied basic research scheme (No: 2024A04J3367), and the NSF of China (No: 62002090).

References

Buch, S.; Eyzaguirre, C.; Gaidon, A.; Wu, J.; Fei-Fei, L.; and Niebles, J. C. 2022. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2917–2927.

Cong, Y.; Liao, W.; Ackermann, H.; Rosenhahn, B.; and Yang, M. Y. 2021. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 16372–16382.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9268–9277.

Feng, S.; Mostafa, H.; Nassar, M.; Majumdar, S.; and Tripathi, S. 2023. Exploiting long-term dependencies for generating dynamic scene graphs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5130–5139.

Gao, D.; Zhou, L.; Ji, L.; Zhu, L.; Yang, Y.; and Shou, M. Z. 2023a. MIST: Multi-modal Iterative Spatial-Temporal Transformer for Long-form Video Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14773–14783.

Gao, K.; Chen, L.; Huang, Y.; and Xiao, J. 2021. Video relation detection via tracklet based visual transformer. In *Proceedings of the 29th ACM international conference on multimedia*, 4833–4837.

Gao, K.; Chen, L.; Niu, Y.; Shao, J.; and Xiao, J. 2022. Classification-then-grounding: Reformulating video scene graphs as temporal bipartite graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19497–19506.

Gao, K.; Chen, L.; Zhang, H.; Xiao, J.; and Sun, Q. 2023b. Compositional prompt tuning with motion cues for open-vocabulary video relation detection. *arXiv preprint arXiv:2302.00268*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Ji, J.; Krishna, R.; Fei-Fei, L.; and Niebles, J. C. 2020. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10236–10247.

Johnson, J.; Krishna, R.; Stark, M.; Li, L.-J.; Shamma, D.; Bernstein, M.; and Fei-Fei, L. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3668–3678.

Li, R.; Zhang, S.; Wan, B.; and He, X. 2021. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11109–11119.

Li, Y.; Ouyang, W.; Zhou, B.; Wang, K.; and Wang, X. 2017. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE international conference on computer vision*, 1261–1270.

Li, Y.; Yang, X.; and Xu, C. 2022. Dynamic scene graph generation via anticipatory pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13874–13883.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, X.; Ding, C.; Zeng, J.; and Tao, D. 2020. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3746–3753.

Lin, X.; Ding, C.; Zhan, Y.; Li, Z.; and Tao, D. 2022a. Hl-net: Heterophily learning network for scene graph generation. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19476–19485.

Lin, X.; Ding, C.; Zhang, J.; Zhan, Y.; and Tao, D. 2022b. Ru-net: Regularized unrolling network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19457–19466.

Liu, C.; Jin, Y.; Xu, K.; Gong, G.; and Mu, Y. 2020. 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*, 10840–10849.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, C.; Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2016. Visual relationship detection with language priors. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 852–869. Springer.

Nag, S.; Min, K.; Tripathi, S.; and Roy-Chowdhury, A. K. 2023. Unbiased Scene Graph Generation in Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22803–22813.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Shang, X.; Di, D.; Xiao, J.; Cao, Y.; Yang, X.; and Chua, T.-S. 2019. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, 279–287.

Shang, X.; Ren, T.; Guo, J.; Zhang, H.; and Chua, T.-S. 2017. Video visual relation detection. In *Proceedings of the 25th ACM international conference on Multimedia*, 1300–1308.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer*

Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, 510–526. Springer.

Tang, K.; Zhang, H.; Wu, B.; Luo, W.; and Liu, W. 2019. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6619–6628.

Teng, Y.; Wang, L.; Li, Z.; and Wu, G. 2021. Target adaptive context aggregation for video scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13688–13697.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, S.; Gao, L.; Lyu, X.; Guo, Y.; Zeng, P.; and Song, J. 2022. Dynamic scene graph generation via temporal prior inference. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5793–5801.

Xiao, J.; Zhou, P.; Chua, T.-S.; and Yan, S. 2022. Video graph transformer for video question answering. In *European Conference on Computer Vision*, 39–58. Springer.

Xie, Y.; Dai, H.; Chen, M.; Dai, B.; Zhao, T.; Zha, H.; Wei, W.; and Pfister, T. 2020. Differentiable top-k with optimal transport. *Advances in Neural Information Processing Systems*, 33: 20520–20531.

Yang, X.; Gao, C.; Zhang, H.; and Cai, J. 2020. Hierarchical scene graph encoder-decoder for image paragraph captioning. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4181–4189.

Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5831–5840.

Zhang, J.; Shih, K. J.; Elgammal, A.; Tao, A.; and Catanzaro, B. 2019. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11535–11543.

Zheng, C.; Lyu, X.; Gao, L.; Dai, B.; and Song, J. 2023. Prototype-based Embedding Network for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22783–22792.